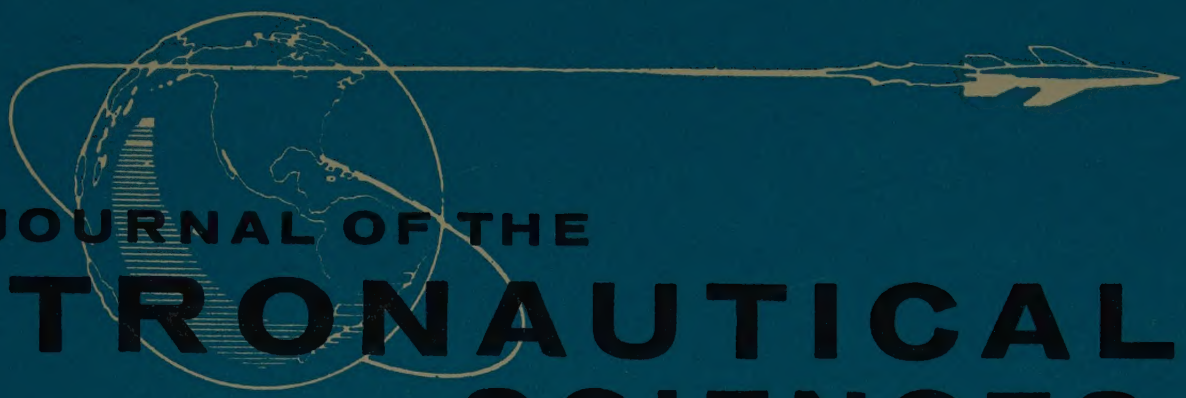

THE JOURNAL OF THE ASTRONAUTICAL SCIENCES

VOLUME VIII, NO. 4 WINTER 1961

CONTENTS

PUBLISHED BY THE AMERICAN ASTRONAUTICAL SOCIETY, INC.

THE AMERICAN ASTRONAUTICAL SOCIETY, INC.

516 Fifth Avenue, New York 36, New York, U.S.A.

THE AMERICAN ASTRONAUTICAL SOCIETY

The American Astronautical Society, founded in 1953 and incorporated in New York State in 1954, is a national scientific organization dedicated to advancement of the astronautical sciences. The society considers manned interplanetary space flight a logical progression from today's high-performance research aircraft, guided missile, and earth satellite operations. The scope of the society is illustrated by a partial list of the astronautical fields of interest: astronavigation, biochemistry, celestial mechanics, cosmology, geophysics, space medicine, and upper atmosphere physics, as well as the disciplines of astronautical engineering, including space vehicle design, communications, control, instrumentation, guidance, and propulsion. The aims of the society are to encourage scientific research in all fields related to astronautics and to propagate knowledge of current advances. Promotion of astronautics in this way is accomplished by the society largely through its program of technical meetings and publications

AFFILIATIONS

AAS cooperates with other national and international scientific and engineering organizations. AAS is an affiliate of the American Association for the Advancement of Science and a member organization of the International Astronautical Federation.

MEMBERSHIP REQUIREMENTS

All persons having a sincere interest in astronautics or engaged in the practice of any branch of science, which contributes to or advances the astronautical sciences, are eligible for one of the various grades of membership in the Society. Requirements are tabulated below. A special category of Student Membership has been authorized for full time students or those under 18 years of age. A nominal membership fee of $5.00 is made in such cases to cover publications. The Directors of the Society may elect as Fellows of the Society those who have made direct and significant contributions to the astronautical sciences. Information regarding individual membership as well as Corporate and Benefactor Membership may be obtained by writing the Corresponding Secretary at the Society address.

Grade	Contribution To Astronautics	Experience or Scientific Training*	Annual Dues
Affiliate Member	Interest	none required	$8
Member	Active Interest	8 years	$10
Senior Member	Recognized Standing and Direct Contribution	10 years	$15

* A Bachelor's, Master's or Doctor's degree in any branch of science or engineering is equivalent to four, six or eight years of experience, respectively.

The *Journal of the Astronautical Sciences*

Director of Publications, George R. Arthur
Editor, Robert M. L. Baker, Jr.
Managing Editor, George J. Clark

Published quarterly by the AMERICAN ASTRONAUTICAL SOCIETY, INC. at 428 E. Preston Street, Baltimore 2, Maryland

Address all Journal correspondence to Box 24721, Los Angeles 24, Calif.

Subscription Rates: One year $5.00; foreign $6.00; single copy $1.25. The Journal is published quarterly.

Second-class postage paid at Baltimore, Maryland.

Green's Functions for Space Trajectory Perturbation Analysis[1]

Joseph C. Dunn[2]

Abstract

A first order perturbation analysis is developed for the restricted n-body problem of interplanetary flight. The linearized perturbation equations, derived from a generalized functional representation of the first and second integrals of the dynamical differential equations, appear in a modified tensor notation (without recourse to the summation convention) which underscores their inherent structural simplicity. At the heart of this formulation is a time dependent Green's function, or influence tensor, which is derived for both fixed terminal times (upper limits of integration) and open terminal times defined by some terminal constraint. The text concludes with a more or less complete mathematical development of two techniques for evaluating the fixed terminal time Green's function. Direct integration, which constitutes the first method, is at once the most straightforward and obvious of the pair. It is included here simply for contrast with the second method, an indirect form of "integration" that utilizes the properties of a system of equations adjoint to a linearized form of the trajectory differential equations. A self-contained mathematical derivation of the latter technique appears, for convenience, in an appended discussion.

Introduction

Of late, analyses designed to predict the optimum magnitude and spacing of corrective thrust application for maneuverable space vehicles have solicited considerable interest. At the same time, the companion problem of trajectory and orbit determination in the face of measurement errors inherent in available tracking systems deserves, and has received, attention. A common requirement for both types of studies is the capability to predict a vehicle's departure from a so-called precalculated nominal or standard trajectory at a specified point in time, caused by position and/or velocity increments introduced at a previous time. This capability is most generally achieved through a mathematical formulation based upon certain linearity assumptions which require only that the perturbations considered (be they actual velocity increments caused by thrust impulses, or position and/or velocity measurement errors) are small quantities. The linearized approach ultimately involves the calculation of a Green's function whose components are sometimes referred to as error coefficients.

[1] Manuscript received December 2, 1960, revised August 4, 1961.

[2] Research Engineer, Grumman Aircraft Engineering Corporation, Bethpage, New York.

The primary objective of this paper is to put before the analyst a brief but unified theoretical presentation of linearized perturbation theory as it applies to interplanetary trajectory studies. An attempt has been made to place the emphasis, as much as is possible, on mathematical rigor in order to complement the expositions of practical application which have already been made in some detail by Battin, Breakwell, Laning, and Pfeiffer, [1], [2], [3], [4], and [5]. Where the need arises, a few simple examples have been inserted for clarification.

Linearized Trajectory Perturbation Equations for the Restricted Three-Dimensional n-Body Problem

The differential equations describing the motion of several mutually attractive central gravitational force fields which are in transit with respect to an inertial space,[3] may be written as follows [6]:

$$\ddot{x}_i = -K \sum_{j=1}^{n} m_j \frac{(x_i - x_j)}{r_{ij}^3}, \tag{1a}$$

$$\ddot{y}_i = -K \sum_{j=1}^{n} m_j \frac{(y_i - y_j)}{r_{ij}^3}, \quad \begin{matrix} i = 1,n \\ j = 1,n; j \neq i \end{matrix} \tag{1b}$$

$$\ddot{z}_i = -K \sum_{j=1}^{n} m_j \frac{(z_i - z_j)}{r_{ij}^3}, \tag{1c}$$

$$r_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2}, \tag{1d}$$

where x_i, y_i, and z_i define the position of the ith mass point relative to an inertial cartesian space, and the superscripted dot indicates a single differentiation in time.

Equations (1a), (1b), and (1c) may be expressed in a symbolic form which simply states the functional dependence, e.g.:

$$\ddot{x}_i = f_i(x_j, y_j, z_j), \tag{2a}$$

$$\ddot{y}_i = g_i(x_j, y_j, z_j), \qquad i = 1, n; j = 1, n \tag{2b}$$

and

$$\ddot{z}_i = h_i(x_j, y_j, z_j). \tag{2c}$$

At time $t = \tau$, define a set of nominal initial conditions in a position-velocity configuration space such that:

$$q_{jl} = q_{jl}(\tau); \qquad l = 1, 6; j = 1, n,$$

[3] General n-body problem of celestial mechanics.

where $q_{j1} = \dot{x}_j$, $q_{j2} = \dot{y}_j$, $q_{j3} = \dot{z}_j$, $q_{j4} = x_j$, $q_{j5} = y_j$, $q_{j6} = z_j$. Then the corresponding first and second integrals of Eqs. (2a), (2b), (2c) are seen to depend solely upon these initial conditions and a nominal terminal time, t_f, (upper limit of integration) counted from τ, i.e.,

$$q_{ik}(t_f) = F_{ik}(q_{jl}(\tau), t_f); \qquad k = 1, 6; i = 1, n. \tag{3}$$

A question presents itself: What increments, Δq_{ik}, in the nominal variables $q_{ik}(t_f)$ are produced by the perturbations $\Delta q_{jl}(\tau)$ and Δt_f, introduced at time $\tau < t_f$? Or, from a physical point of view, how will a number of small variations in individual component orientations and velocities introduced at time τ, ultimately affect the subsequent n-body configuration at a new terminal time $\bar{t}_f = t_f + \Delta t_f$? The answer, of course, can be arrived at by simply reintegrating the n-body equations of motion for the new set of initial conditions $q_{jl}(\tau) + \Delta q_{jl}(\tau)$ and extending the upper limit of integration to $t_f + \Delta t_f$. But, if the perturbations $\Delta q_{jl}(\tau)$ and Δt_f are constrained to remain "small," a more convenient approach to the problem is justified.

By way of definition:

$$\Delta q_{ik}(\bar{t}_f) = \bar{q}_{ik}(\bar{t}_f) - q_{ik}(t_f), \tag{4}$$

where the $q_{ik}(t_f)$ are nominal (unperturbed) parameters evaluated at the nominal terminal time t_f and the $\bar{q}_{ik}(\bar{t}_f)$ are off-nominal parameters evaluated at new terminal time $\bar{t}_f = t_f + \Delta t_f$, following the introduction of $\Delta q_{jl}(\tau)$ perturbations at time τ. The $\bar{q}_{ik}(\bar{t}_f)$ may be estimated, in first approximation, by means of a truncated Taylor series expansion of Eq. (3) in the neighborhood of $q_{jl}(\tau)$ and t_f. Then, to the exclusion of second and higher order terms;

$$\bar{q}_{ik}(\bar{t}_f) = q_{ik}(t_f) + \sum_{j=1}^{n}\sum_{l=1}^{6}\frac{\partial F_{ik}}{\partial q_{jl}}(\tau)\Delta q_{jl}(\tau) + \frac{\partial F_{ik}}{\partial t}(t_f)\,\Delta t_f, \tag{5}$$

assuming that the derivatives $\frac{\partial F_{ik}}{\partial q_{jl}}$ and $\frac{\partial F_{ik}}{\partial t}$ exist and are continuous in the neighborhood of the expansion. With the aid of Eqs. (4) and (5) one finds that:

$$\Delta q_{ik}(\bar{t}_f) = \sum_{j=1}^{n}\sum_{l=1}^{6}\frac{\partial F_{ik}}{\partial q_{jl}}(\tau)\Delta q_{jl}(\tau) + \frac{\partial F_{ik}}{\partial t}(t_f)\Delta t_f. \tag{6}$$

Also, consideration of Eq. (3) reveals that $\frac{\partial F_{ik}}{\partial t}(t_f) = \frac{d}{dt}q_{ik}(t_f) = \dot{q}_{ik}(t_f)$ since;

$$\frac{d}{dt}\bar{q}_{ik}(t_f) = \dot{q}_{ik}(t_f) = \sum_{j=1}^{n}\sum_{l=1}^{6}\frac{\partial F_{ik}}{\partial q_{jl}}(\tau)\frac{d}{dt}q_{jl}(\tau) + \frac{\partial F_{ik}}{dt}(t_f) \tag{7}$$

and;

$$\frac{d}{dt}q_{jl}(\tau) \equiv 0; \qquad \text{(the initial conditions are not functions of time).} \tag{8}$$

Equation (6) can now be written in a somewhat more descriptive notation, viz:

$$\Delta q_{ik}(\bar{t}_f) = \sum_{j=1}^{n}\sum_{l=1}^{6}\frac{\partial F_{ik}}{\partial q_{jl}}(\tau)\Delta q_{jl}(\tau) + \dot{q}_{ik}(t_f)\Delta t_f \tag{9}$$

which comprises the set of $6n$ linearized perturbation equations in $6n$ unknowns, $\Delta q_{ik}(\bar{t}_f)$, for the general n-body problem. Now consider its special application to a restricted n-body problem.

Suppose that only the mth particle of all the constituent elements in an n-body group is subject to perturbations, i.e.;

$$\Delta q_{jl}(\tau) = \delta_{jm}\Delta q_{jl}(\tau); \quad \delta_{jm} = \begin{cases} 0, & j \neq m \\ 1, & j = m \end{cases}$$

Then Eq. (9) simplifies to:

$$\Delta q_{ik}(\bar{t}_f) = \sum_{l=1}^{6}\frac{\partial F_{ik}}{\partial q_{ml}}(\tau)\Delta q_{ml}(\tau) + \dot{q}_{ik}(t_f)\Delta t_f. \tag{10}$$

If it is further assumed that the same mth particle exerts an insignificant influence on the remaining $n - 1$ bodies, then it follows that:

$$\frac{\partial F_{ik}}{\partial q_{ml}}(\tau) = \delta_{im}\frac{\partial F_{ik}}{\partial q_{ml}}(\tau); \quad \delta_{im} = \begin{cases} 0, & i \neq m \\ 1, & i = m \end{cases}$$

Consequently;

$$\Delta q_{mk}(\bar{t}_f) = \sum_{l=1}^{6}\frac{\partial F_{mk}}{\partial q_{ml}}(\tau)\Delta q_{ml}(\tau) + \dot{q}_{mk}(t_f)\Delta t_f \tag{11a}$$

and,

$$\Delta q_{ik}(\bar{t}_f) = \dot{q}_{ik}(t_f)\Delta t_f; \qquad i \neq m. \tag{11b}$$

It is significant that the assumptions which make Eqs. (11a) and (11b) a special case of Eq. (9) have real physical validity for interplanetary trajectory perturbation mechanics. Most certainly, a spacecraft exerts a totally negligible influence on planetary and/or stellar elements in its vicinity: also, one commonly assumes a perfect knowledge of planetary kinematics and dynamics when computing trajectories. Consequently, in this light, the spacecraft is the only member of the n body group (planets + Sun + spacecraft) which is not exempt from perturbative influence. It then must follow that changes in a spacecraft's "state" (position and velocity) at a terminal point may only be attributed to deviations from its initial state at a previous time. At this point, Eqs. (11a) are immediately applicable provided that the deviations from the standard initial state are "small."

In actuality, planetary kinematics and dynamics are not known precisely, and the uncertainties which do exist (e.g., in the astronomical constants, etc.) can result in a definite dispersion from the precalculated terminal state even if initial conditions are met exactly.

An assessment of the influence of uncertainties of this sort can be made by elevating the parameters in doubt to the level of a state variable, and then proceeding to construct a set of expanded perturbation equations. The procedure is identical to that described in earlier pages and, for example, might start from a set of differential equations of the following sort:

$$\ddot{x}_i = f_i(x_j, K_l) \qquad j = 1, n; l = 1, m,$$
$$\dot{K}_l = 0$$

etc., where the K_l's are here taken to be coefficients in the dynamical equations whose values are in doubt.

Unfortunately, further exploration of this trend of thought, however fruitful, does not lie within the protracted scope of this paper: subsequent developments will therefore begin with Eq. (11a) and the implications contained therein.

Henceforth, the subscript m in Eq. (11a) will be deleted and it is to be understood that the parameters, q, are to be associated only with vehicle coordinates and coordinate velocities. Then Eq. (11a) goes over to:

$$\Delta q_k(\bar{t}_f) = \sum_{l=1}^{6} \frac{\partial F_k}{\partial q_l}(\tau)\Delta q_l(\tau) + \dot{q}_k(t_f)\Delta t_f \text{ ;}^{4} \quad (12)$$

where, in summary: $\Delta q_k(\bar{t}_f)$ is a vector whose components are incremental displacement and velocity deviations, at time $\bar{t}_f = t_f + \Delta t_f$, from nominal trajectory variables at time t_f ; $\Delta q_l(\tau)$ is a vector whose components are displacement and velocity perturbations introduced at time $\tau < t_f$; $\dot{q}_k(t_f)$ are nominal accelerations and velocities at time t_f ; and $\frac{\partial F_k}{\partial q_l}(\tau)$ is a Green's function or "influence" function which plays a decisive part in the theory of linearized trajectory perturbation mechanics [5], [7], [8], and [9].

Fixed Terminal Times and the Green's Function

In the next section of this discussion, *Varying Terminal Time Defined by an External Constraint*, the purposes served by the Δt_f perturbation in Eq. (12) will become clear. However, in order to emphasize the significance of the Green's function in linearized perturbation theory, it will be assumed for the moment that the terminal time is fixed at t_f for both nominal and off-nominal trajectories, i.e., $\Delta t_f = 0$. Then Eq. (12) reduces to:

$$\Delta q_k(t_f) = \sum_{l=1}^{6} \frac{\partial F_k}{\partial q_l}(\tau)\Delta q_l(\tau). \quad (13)$$

The Green's function $\frac{\partial F_k}{\partial q_l}(\tau)$, is therefore a second order tensor operator which transforms the perturbation vector, $\Delta q_l(\tau)$, to a new vector whose components are parameter deviations, at the *nominal* terminal time t_f, from a standard trajectory. The partial differential coefficients which are the components of the influence tensor, may be termed, in the most general sense, fixed terminal time, first order perturbation coefficients. (Sometimes referred to as "error coefficients" in guidance and control studies.) Their evaluation will become a topic of investigation on the following pages. With reference to Eq. (13), expression (12) can now be written in another form, viz:

$$\Delta q_k(\bar{t}_f) = \Delta q_k(t_f) + \dot{q}_k(t_f)\Delta t_f . \quad (14)$$

Varying Terminal Time Defined by an External Constraint—The Modified Green's Function

The purpose served by the perturbation, Δt_f, in the foregoing developments has perhaps been somewhat obscure. It is clear, from the physics of the problem, that Δt_f, unlike the $\Delta q_l(\tau)$'s, does not actually alter a nominal trajectory. The only function of this "perturbation" is to provide a means of comparing q_k variables on the altered trajectory at time

$$\bar{t}_f = t_f + \Delta t_f$$

to their respective nominal trajectory counterparts at time t_f. When trajectory deviations are of interest, not at a fixed terminal time, but rather, at a variable terminal time determined by a certain criterion, a mathematical formulation which makes allowances for a Δt_f becomes eminently useful. The following example will attempt to demonstrate how Eq. (12) may be used to calculate $\Delta q_k(\bar{t}_f)$'s, where the time t_f is now assigned to a particular event (in this instance, arrival at closest planetary approach) which, because of previously introduced disturbances, is displaced in time from t_f by a small increment Δt_f.

Consider a special case where the time t_f corresponds to the point of closest planetary approach on some nominal interplanetary trajectory. Then it is reasonable to suspect that if $\Delta q_l(\tau)$ perturbations are introduced at time $\tau < t_f$, the time t_f will find the vehicle at some point other than closest planetary proximity on the off-nominal flight path. It follows, therefore, that if the terminal time $\bar{t}_f$ of interest is now the time of off-nominal closest approach, then $\bar{t}_f$, in general, differs from t_f by a small increment Δt_f which must be determined by means of the appropriate equation of constraint which must be satisfied at both the nominal and off-nominal times, i.e., $\dot{r}_p(t_f) = \dot{\bar{r}}_p(\bar{t}_f) = 0$.[5]

The transformation equations which relate planet-centered spherical coordinates to inertial coordinates may be written in symbolic form which, for example, expresses the dependence of r_p as

$$r_p(t) = P[x(t), y(t), z(t), t]. \quad (15)$$

The time, t, appears separately here by virtue of the fact that the target planet would, in general, be moving

[4] The reader will note that in appearance and manner of derivation these equations are analogous to the differential correction equations of classical orbital mechanics [10].

[5] Where r_p is the vehicle-target position vector.

in some prescribed fashion against the background of inertial space. From Eq. (15), it follows that:

$$\dot{r}_p(t) = \frac{\partial P}{\partial x}\frac{dx}{dt} + \frac{\partial P}{\partial y}\frac{dy}{dt} + \frac{\partial P}{\partial z}\frac{dz}{dt} + \frac{\partial P}{dt} = Q[x(t), y(t), z(t), \dot{x}(t), \dot{y}(t), \dot{z}(t), t]. \tag{16}$$

Since the parameters x, y, z, etc. are associated with the position of a vehicle, then Eq. (16) may be written in the previously established notation, viz:

$$\dot{r}_p(t_f) = Q[q_s(t_f), t_f] \qquad s = 1, 6.$$

Obviously, if the parameters $q_s(t_f)$ and t_f are subject to small variations, then $\dot{r}_p(t_f)$ must also change. Denote by $\dot{\bar{r}}_p(\bar{t}_f)$, the vehicle's radial velocity on a perturbed trajectory at time $\bar{t}_f = t_f + \Delta t_f$, and, by $\dot{r}_p(t_f)$, the vehicle's radial velocity on a nominal trajectory at time t_f. Then $\dot{\bar{r}}_p(\bar{t}_f)$ and $\dot{r}_p(t_f)$ are related by:

$$\dot{\bar{r}}_p(\bar{t}_f) = \dot{r}_p(t_f) + \Delta\dot{r}_p(\bar{t}_f) = \dot{r}_p(t_f) + \sum_{s=1}^{6} \frac{\partial Q}{\partial q_s} \cdot (t_f)\Delta q_s(t_f) + \frac{\partial Q}{\partial t}(t_f)\Delta t_f. \tag{17}$$

The significance of $\frac{\partial Q}{\partial t}(t_f)$ in Eq. (17) is, in this instance, more forcefully demonstrated by a deductive reasoning process rather than straightforward manipulation of Eq. (16).

Suppose $\Delta t_f = 0$ in Eq. (17); then:

$$\dot{\bar{r}}_p(t_f) = \dot{r}_p(t_f) + \Delta\dot{r}_p(t_f) = \dot{r}_p(t_f) + \sum_{s=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\Delta q_s(t_f). \tag{17a}$$

Equation (17a) determines the off-nominal value of $\dot{r}_p$ at the *nominal* terminal time t_f. In general, an off-nominal radial acceleration, $\ddot{\bar{r}}_p(t_f)$, will also exist at the same instant in time. Then, by linear approximation:

$$\dot{\bar{r}}_p(\bar{t}_f) = \dot{\bar{r}}_p(t_f) + \ddot{\bar{r}}_p(t_f)\Delta t_f = \dot{r}_p(t_f) + \sum_{s=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\Delta q_s(t_f) + \ddot{\bar{r}}_p(t_f)\Delta t_f. \tag{18}$$

However, from Eq. (17a), it follows that:

$$\ddot{\bar{r}}_p(t_f) \equiv \ddot{r}_p(t_f) + \frac{d}{dt}[\Delta\dot{r}_p(t_f)].$$

But, $\frac{d}{dt}$ and Δ are commutative operations in this instance. Therefore:

$$\ddot{\bar{r}}_p(t_f) = \ddot{r}_p(t_f) + \Delta\ddot{r}_p(t_f).$$

Consequently, Eq. (18) reduces to:

$$\dot{\bar{r}}_p(\bar{t}_f) = \dot{r}_p(t_f) + \sum_{s=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\Delta q_s(t_f) + [\ddot{r}_p(t_f) + \Delta\ddot{r}_p(t_f)]\,\Delta t_f. \tag{18a}$$

It is clear that $\ddot{r}_p(t_f)$ is never zero, and $\Delta\ddot{r}_p(t_f)/\ddot{r}_p(t_f)$ will likely be a negligible quantity provided that the initial disturbance is moderate. Therefore, to the first order of small quantities;

$$\dot{\bar{r}}_p(\bar{t}_f) = \dot{r}_p(t_f) + \sum_{s=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\Delta q_s(t_f) + \ddot{r}_p(t_f)\Delta t_f. \tag{18b}$$

Comparison of Eqs. (17) and (18b) leads to the following conclusion:

$$\frac{\partial Q}{\partial t}(t_f) = \ddot{r}_p(t_f) = \dot{Q}(t_f),$$

which parallels that arrived at for $\frac{\partial F_k}{dt}$ in Eq. (12).

Equation (18b) [or its equivalent, Eq. (17)] may now be cast in still another form since an examination of Eqs. (13) reveals that:

$$\Delta q_s(t_f) = \sum_{l=1}^{6} \frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau). \tag{19}$$

Therefore Eq. (18b) goes over to:

$$\dot{\bar{r}}_p(\bar{t}_f) = \dot{r}_p(t_f) + \sum_{s=1}^{6}\sum_{l=1}^{6} \frac{\partial Q}{\partial q_s}(t_f) \cdot \frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau) + \ddot{r}_p(t_f)\Delta t_f. \tag{20}$$

The necessary conditions for r_p to be a minimum on the nominal trajectory at time t_f, and on the off-nominal trajectory at time $\bar{t}_f$, require that $\dot{r}_p(t_f) = 0$ and $\dot{\bar{r}}_p(\bar{t}_f) = 0$.[6] Under these circumstances Eq. (20) reduces to:

$$0 = 0 + \sum_{s=1}^{6}\sum_{l=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau) + \dot{Q}(t_f)\Delta t_f, \tag{21}$$

from which it follows that Δt_f cannot be arbitrarily specified in the presence of a constraint, but rather, is uniquely defined by the vector $\Delta q_l(\tau)$, i.e.:

$$\Delta t_f = -\frac{1}{\dot{Q}(t_f)} \sum_{s=1}^{6}\sum_{l=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau), \tag{21a}$$

which constitutes an intuitively acceptable result, since the displacement in time of closest approach is a direct consequence of the $\Delta q_l(\tau)$ perturbations.

Admitting to Eq. (21a), the small changes $\Delta q_k(\bar{t}_f)$ in the trajectory parameters at the point of off-nominal closest planetary approach must be given by a special form of Eq. (12), viz:

$$\Delta q_k(\bar{t}_f) = \sum_{l=1}^{6} \frac{\partial F_k}{\partial q_l}(\tau)\Delta q_l(\tau) - \frac{\dot{q}_k(t_f)}{\dot{Q}(t_f)} \sum_{s=1}^{6}\sum_{l=1}^{6} \frac{\partial Q}{\partial q_s}(t_f)\frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau) \tag{22}$$

[6] The vanishing of $\dot{r}_p$ is a necessary but not sufficient condition for r_p to assume a minimum value. However, the distinction between maxima and minima can, in this example, be made quite readily through physical reasoning alone; otherwise, the derivative $\ddot{r}_p$, must be taken into consideration.

or

$$\Delta q_k(\bar{t}_f) = \sum_{l=1}^{6}\sum_{s=1}^{6}\left[\delta_{sk} - \frac{\dot{q}_k(t_f)}{\dot{Q}(t_f)}\frac{\partial Q}{\partial q_s}(t_f)\right] \cdot \frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau); \qquad \delta_{sk} = \begin{cases} 0, & s \neq k \\ 1, & s = k. \end{cases} \tag{22a}$$

If $A_{sk}(t_f) = \left[\delta_{sk} - \frac{\dot{q}_k(t_f)}{\dot{Q}(t_f)}\frac{\partial Q}{\partial q_s}(t_f)\right]$, then (22a) can be written in a more compact form given by:

$$\Delta\dot{q}_k(\bar{t}_f) = \sum_{l=1}^{6}\sum_{s=1}^{6} A_{sk}(t_f)\frac{\partial F_s}{\partial q_l}(\tau)\Delta q_l(\tau) = \sum_{l=1}^{6} A'_{kl}\,\Delta q_l(\tau), \tag{22b}$$

where the second order tensor A'_{kl} is simply a modified Green's function which is compatible with the mathematical constraint imposed at t_f and $\bar{t}_f$.

One final word on this subject is in order: Although in the preceding example the mathematical constraint was seen to enforce a stationary value for the vehicle-target position vector, the reader should bear in mind that a modified Green's Function such as A'_{kl} might easily be derived for a variety of terminal constraints. As a matter of record, the tensor $A_{sk}(t_f)$, as it appears in Eq. (22b), is representative of any pertinent mathematical constraint having the form $Q(t_f) = \bar{Q}(\bar{t}_f) = 0$, $(\bar{t}_f = t_f + \Delta t_f)$ and subject to the proviso that $Q = Q[q_s(t), t]$.

Variable "Initial" Times

It has previously been stated that the vector $q_k(t_f)$ is a function of the vector $q_l(\tau)$ and the time t_f counted from τ, i.e.:

$$q_k(t_f) = F_k(q_l(\tau), t_f). \tag{23}$$

However, if τ is, in fact, an intermediate time and the set of six parameters $q_m(0)$ are associated with the time $t = 0 \leqq \tau \leqq t_f$, then:

$$q_l(\tau) = G_l(q_m(0), \tau). \tag{24}$$

Consequently;

$$q_k(t_f) = F_k\{G_l[q_m(0), \tau], t_f\}. \tag{25}$$

Equations (23), (24), and (25), together, are nothing more than a disguised form of the elementary integral relationship which guarantees that:

$$\int_0^{t_f} \dot{q}_k(t)\,dt = \int_0^{\tau} \dot{q}_k(t)\,dt + \int_{\tau}^{t_f} \dot{q}_k(t)\,dt. \tag{26}$$

On the strength of Eq. (26), Eq. (25) can be carried one step further to:

$$q_k(t_f) = F_k'[q_m(0), t_f] \tag{27}$$

and the intermediate time τ completely disappears from the formulation provided that t_f is now counted from $t = 0$. In this context, τ retains its initial character only in the sense that a perturbation vector $\Delta q_l(\tau)$ is introduced at this point in time.

An examination of the tensor A_{sk} in Eq. (22b) discloses the dependence of its components on the nominal quantities $\dot{q}_k$, $\ddot{r}_p$, and $\frac{\partial Q}{\partial q_s}$, evaluated at the nominal terminal time t_f. But $\ddot{r}_p$ and the derivatives $\frac{\partial Q}{\partial q_s}$ are themselves, readily derivable functions of q_k and $\dot{q}_k$, thus one is led to the rather remarkable result that the tensor A_{sk} is uniquely defined by the standard trajectory's initial conditions, its nominal terminal time, t_f, and the nature of the terminal constraint, *without consideration for the time of perturbation, τ or the perturbation vector itself.* On the other hand, the magnitudes of the partial differential coefficients, which are the components of the Green's function $\frac{\partial F_k}{\partial q_l}(\tau)$, are related to the size of the contribution made by the second integral on the right-hand side of Eq. (26) since the $\Delta q_l(\tau)$ perturbation vector must exert its influence on $q_k(t_f)$ through this integral. *Consequently, as one might expect, these coefficients and the associated influence tensor are functions of the running variable τ;* $0 \leqq \tau \leqq t_f$. The situation is further aggravated by the fact that the functions F_k are, in general, not available in closed form, therefore, the components of $\frac{\partial F_k}{\partial q_l}(\tau)$ must be evaluated by a relatively laborious numerical procedure. However, in this instance, the end more than justifies the means, since a time history of the Green's function (along with the invariant A_{sk} tensor) completely describes the sensitivity of the $q_k(\bar{t}_f)$ variables to any *arbitrary* (small) perturbation vector introduced at any intermediate time τ on a nominal trajectory. In fact, if care is taken to insure against a violation of the small disturbance assumption, the principle of superposition holds and the perturbation analysis can be extended to include situations where the components of $\Delta q_l(\tau)$ are continuous (or piecewise continuous) functions of time on an interval $\tau_2 - \tau_1$, e.g.: let $S\Delta q_k(\bar{t}_f)$ denote the *total* increments in $q_k(t_f)$, then:

$$S\Delta q_k(\bar{t}_f) = \int_{\tau_1}^{\tau_2}\sum_{l=1}^{6}\sum_{s=1}^{6} A_{sk}\frac{\partial F_s}{\partial q_l}(t)\,dq_l(t) = \sum_{l=1}^{6}\sum_{s=1}^{6} A_{sk}\int_{\tau_1}^{\tau_2}\frac{\partial F_s}{\partial q_l}(t)\,dq_l(t). \tag{28}$$

As an illustration of a situation where Eq. 28 is physically meaningful, one might reflect on a corrective guidance scheme which employs relatively low levels of thrust for finite periods of time (as opposed to impulsive corrections). Let $u_i(t)$, $i = 1, 2, 3$, represent the components of the imparted acceleration vector, $\bar{T}/m$, where m is the mass of the spacecraft and $\bar{T}$ is the corrective thrust vector. Then;

$$d\dot{x}_i(t) = u_i(t)\,dt,$$

and therefore, the total changes in terminal state affected by the control function are given by integrals of the form,

$$\sum_{l=1}^{3} \int_{\tau_1}^{\tau_2} \frac{\partial F_s}{\partial(d\dot{x}_l)}(t) u_l(t)\, dt,$$

where $[\tau_1, \tau_2]$ represents the interval during which the corrective thrust is applied.

In any event, the evaluation of the Green's Function $\frac{\partial F}{\partial q_l}(\tau)$ is, at once, the essence and central problem of linearized perturbation mechanics. Attention is now directed toward two possible operational techniques which will generate the influence tensor.

Evaluation of the Green's Function by Direct and Indirect Integration

From a conceptual standpoint, the simplest and most straightforward approach to be followed in evaluating the influence tensor is direct integration. Examine one of the thirty-six components of the Green's function, e.g., $\frac{\partial F_1}{\partial q_1}(\tau)$. By definition:

$$\frac{\partial F_1}{\partial q_1}(\tau) = \lim_{\Delta q_1(\tau) \to 0} \{F_1[q_1(\tau) + \Delta q_1(\tau), q_2(\tau), q_3(\tau), \cdots, t_f] - F_1[q_1(\tau), q_2(\tau), q_3(\tau), \cdots, t_f]\}/\Delta q_1(\tau) \tag{29}$$

or if $\Delta q_1(\tau)$ is sufficiently small;

$$\frac{\partial F_1}{\partial q_1}(\tau) \simeq \{F_1[q_1(\tau) + \Delta q_1(\tau), q_2(\tau), q_3(\tau), \cdots, t_f] - F_1[q_1(\tau), q_2(\tau), q_3(\tau), \cdots, t_f]\}/\Delta q_1(\tau). \tag{29a}$$

The second term in the numerator of Eq. (29a) is, with reference to Eq. (23) and the definitions of the q variables, the nominal x component of the vehicle's velocity at the nominal terminal time t_f, and therefore is well known when the standard trajectory has been selected. The first term can be evaluated by means of direct simultaneous numerical integrations of the equations of motion, which, for the restricted n-body problem, take the form

$$\ddot{x} = f(x, y, z, t), \tag{30a}$$

$$\ddot{y} = g(x, y, z, t), \tag{30b}$$

and

$$\ddot{z} = h(x, y, z, t), \tag{30c}$$

where, once again, t appears separately to allow for planetary movements.

The integrations proceed from the time τ at which the initial conditions $q_1(\tau) + \Delta q_1(\tau)$, $\Delta q_2(\tau)$, etc. are impressed, to the nominal final time t_f. Here $q_1(\tau)$, $q_2(\tau)$, etc. are nominal intermediate parameter values available from the standard trajectory at time τ and $\Delta q_1(\tau)$ is a small but otherwise arbitrarily selected variance. When the integrations have been accomplished, the difference quotient in Eq. (29a) is completely defined and the degree of accuracy of its approximation to the partial differential coefficient $\frac{\partial F_1}{\partial q_1}(\tau)$ is ideally a simple function of the size of the $\Delta q_1(\tau)$ variance; the smaller the variance, the higher the precision. In a similar manner, the remaining thirty-five components of $\frac{\partial F_k}{\partial q_l}(\tau)$ can be obtained.

The direct integration method suffers from one major drawback; the entire procedure outlined above must be repeated each time the lower limit of integration, τ is changed. Obviously, the difficulty is caused by the varying values of $q_l(\tau)$ and the changing interval of integration, $t_f - \tau$, as τ ranges from $t = 0$ to $t = t_f$. A complete time history of the Green's function, as generated by this technique, would therefore require a discouraging amount of numerical labor. Fortunately, another avenue of approach is available which achieves the same results as direct integration with a considerably greater economy in computational steps. This alternative method, which might be described as an indirect integration, relies on the properties of the adjoint to a system of simultaneous linear ordinary differential equations and has been previously applied to problems of this kind by Pfeiffer [5], and to a related problem by Kelley, [7]. Supplementary discussions may be found in Bliss, Goursat, and Goodman and Lance, [9], [11], and [12]. However, in the interest of maintaining the self-sufficiency of this discussion, a development of the necessary mathematics is included in an appendix to this discussion.

The trajectory Eqs. (30a), (30b), and (30c) take the following form when linearized about a nominal path:

$$\Delta\ddot{x} \equiv \frac{d}{dt}(\Delta\dot{x}) = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z, \tag{31a}$$

$$\Delta\ddot{y} \equiv \frac{d}{dt}(\Delta\dot{y}) = \frac{\partial g}{\partial x}\Delta x + \frac{\partial g}{\partial y}\Delta y + \frac{\partial g}{\partial z}\Delta z; \quad (\Delta t = 0), \tag{31b}$$

and

$$\Delta\ddot{z} \equiv \frac{d}{dt}(\Delta\dot{z}) = \frac{\partial h}{\partial x}\Delta x + \frac{\partial h}{\partial y}\Delta y + \frac{\partial h}{\partial z}\Delta z. \tag{31c}$$

These three second order linear ordinary differential equations can be expanded to a set of six first order differential equations by employing the original definition of the q variables, and the following identities:

$$\frac{d}{dt}(\Delta x) \equiv \Delta\dot{x}, \tag{32a}$$

$$\frac{d}{dt}(\Delta y) \equiv \Delta\dot{y}, \tag{32b}$$

and

$$\frac{d}{dt}(\Delta z) \equiv \Delta \dot{z}, \tag{32c}$$

then, Eqs. (31a), (31b), and (31c) can be written in a normal form, viz:

$$\frac{d}{dt}(\Delta q_k) = \sum_{l=1}^{6} a_{kl}\,\Delta q_l\,, \tag{33}$$

where

$$a_{kl} = \begin{bmatrix} 0 & 0 & 0 & \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} & \frac{\partial f}{\partial z} \\ 0 & 0 & 0 & \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} & \frac{\partial g}{\partial z} \\ 0 & 0 & 0 & \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} & \frac{\partial h}{\partial z} \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

the derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ etc., which are the time dependent coefficients of the linear equations, are evaluated along the nominal path.

The complete solution of Eq. (33) may be written in the following form:

$$\Delta q_k(t_f) = \sum_{l=1}^{6} A_{kl}\Delta q_l(\tau). \tag{34}$$

At this point, it should be apparent that A_{kl} above is no more than the influence function of the preceding discussion. Furthermore, in accord with the results developed in Appendix A, a complete time history of A_{kl}, as τ ranges from 0 to t_f, can be generated by a backwards integration of the system of differential equations adjoint to Eq. (33). The adjoint equations are, in this instance;

$$\frac{d}{dt}\lambda_k = -\sum_{l=1}^{6} b_{kl}\,\lambda_l\,; \qquad b_{kl} = a_{lk}\,. \tag{35}$$

The solution vectors, $\lambda_k(\tau)$, $(t_f \leqq \tau \leqq 0)$ which correspond in turn to:

$$\lambda_k(t_f) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \lambda_k(t_f) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ etc.}$$

are identically equal to the row vectors of the tensor A_{kl}.[7] Thus the arduous process of direct integration, with sequential reinitializations is not really required: six uninterrupted backward integrations of the adjoint equations with proper initial conditions imposed once and for all at t_f achieve the identical result.

[7] See Appendix A.

Conclusion

For the sake of brevity and generality, references to specific problems, which would necessitate explicit rather than functional equations, have been purposely avoided. In certain instances, the reader may have found this lack of specialized illustration an aid rather than a hindrance to understanding. The author does regret, however, that the protracted scope of this paper did not allow for a decidedly more complete discussion of magnitude limits for the perturbation vector components. The question of how small is a small perturbation (within the framework of the first order assumptions) has not been adequately considered and, in all probability, cannot be completely answered outside of a specific problem reference frame. Consequently, even though the function and operation of the linearized mechanics has been made reasonably clear in the main body of the text, the allowable extent of its application is still somewhat vague.

Apart from all this, the amount of space devoted to the explanation of indirect integration deserves one last comment. When one appreciates the substantial amount of computing time required to derive a complete time history of the Green's function by direct integration, the savings afforded by the indirect technique are significant. Thus, from the point of view of economy alone, to say nothing of sophistication, indirect integration is the more desirable approach. For this reason, the emphasis has been placed on the associated mathematics in order to provide a firm foundation for the application of this technique to specific problem areas.

Appendix A

Adjoint Differential Equations

Consider the determinate set of n simultaneous, linear, ordinary differential equations in n dependent variables α_i given by:

$$\frac{d}{dt}\alpha_i = \sum_{j=1}^{n} a_{ij}\,\alpha_j\,; \qquad i = 1, n \tag{A1}$$

where the a_{ij}'s are real functions of the independent variable t (which, for the purposes of this discussion, is time) and solutions are desired for the α_i's.

The system of equations adjoint to Eq. (A1) may be written in the following form.[8]

$$\frac{d}{dt}(\lambda_i) = -\sum_{j=1}^{n} b_{ij}\,\lambda_j\,; \qquad i = 1, n, \tag{A2}$$

where

$$b_{ij} = a_{ji}\,. \tag{A2a}$$

[8] For the moment, the physical significance of the λ variables is undefined.

It will now be shown that the integrals of the basic set of differential equations and the integrals of the adjoint set are interrelated in a very useful fashion.

Since α_j and λ_j are characteristically vectors in a parameter space, then their inner product is defined in the usual manner, viz:

$$\lambda \cdot \alpha = \sum_{i=1}^{n} \lambda_i \alpha_i \tag{A3}$$

and the first derivative of this scalar function with respect to the independent variable t is:

$$\frac{d}{dt}(\lambda \cdot \alpha) = \sum_{i=1}^{n} \lambda_i \frac{d}{dt} \alpha_i + \sum_{i=1}^{n} \alpha_i \frac{d}{dt} \lambda_i \tag{A4}$$

which, with the aid of Eqs. (A1) and (A2) reduces to:

$$\frac{d}{dt}(\lambda \cdot \alpha) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i a_{ij} \alpha_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i b_{ij} \lambda_j . \tag{A5}$$

But i and j are dummy indices of summation. They therefore may be interchanged in the second sum of Eq. (A5) to give:

$$\frac{d}{dt}(\lambda \cdot \alpha) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i a_{ij} \alpha_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_j b_{ji} \lambda_i . \tag{A6}$$

But $b_{ji} = a_{ij}$, therefore

$$\frac{d}{dt}(\lambda \cdot \alpha) = 0. \tag{A7}$$

Equation (A7) leads immediately to the interesting conclusion that the inner product of the α vector and the λ vector is invariant with time. Then, if the vectors $\alpha(t_f)$ and $\lambda(t_f)$ are associated with a specific value of the independent variable, it follows that:

$$\sum_{i=1}^{n} \lambda_i(\tau)\alpha_i(\tau) = \sum_{i=1}^{n} \lambda_i(t_f)\alpha_i(t_f), \tag{A8}$$

where τ is any arbitrary value of t different from t_f.

Suppose that in the *adjoint system*, the independent variable increases from t_f in the direction of τ, i.e., $t_f < \tau$. Then the $\lambda_i(\tau)$'s may be obtained by a simultaneous integration of the Eqs. (A2) when a set of initial conditions $\lambda_i(t_f)$ for the adjoint variables has been specified. Let $\lambda_i^{(m)}$ denote a *particular* set of initial adjoint variables. Then Eq. (A8) adjusts itself to:

$$\sum_{i=1}^{n} \lambda_i^{(m)}(\tau)\alpha_i(\tau) = \sum_{i=1}^{n} \lambda_i^{(m)}(t_f)\alpha_i(t_f), \tag{A9}$$

where the $\lambda_i^{(m)}(\tau)$ form the set of n adjoint variables at $t = \tau$ which corresponds to an mth set of initial conditions $\lambda_i^{(m)}(t_f)$. In this context, $\lambda_i^{(m)}(\tau)$ and $\lambda_i^{(m)}(t_f)$ are still vector quantities. But if n *independent* but otherwise arbitrary sets of initial conditions, $\lambda_i^{(m)}(t_f)$, are considered as m ranges from 1 to n, then clearly, $\lambda_i^{(m)}(\tau)$ and $\lambda_i^{(m)}(t_f)$ can be completely contained by the second order tensors, $\lambda_{mi}(\tau)$ and $\lambda_{mi}(t_f)$.

Then Eq. (A9) goes over to a *set* of n simultaneous *independent* algebraic equations in $\alpha_i(\tau)$ and $\alpha_i(t_f)$, viz:

$$\sum_{i=1}^{n} \lambda_{mi}(\tau)\alpha_i(\tau) = \sum_{i=1}^{n} \lambda_{mi}(t_f)\alpha_i(t_f); \quad m = 1, n. \tag{A10}$$

Now, if the flow of time for the α variables is the exact reverse of that for the λ variables, i.e., if $\tau < t_f$, then the $\alpha_i(\tau)$'s are, in effect, initial conditions impressed at a varying lower limit of integration τ. However, the $\alpha_i(t_f)$'s which correspond to these initial conditions may now be obtained by utilizing Eqs. (A10) rather than by directly integrating the α Eqs. (A1). The advantage of this approach comes directly from the fact that, because of the purposeful time flow reversal, the *lower* limit of integration, τ, for the α system becomes the upper limit of integration for the $\lambda_{mi}(\tau)$ variables *which, together with the associated* $\lambda_{mi}(t_f)$, *and* $\alpha_i(\tau)$ *are the coefficients of Eqs. (A10).*

Consequently, the $\lambda_{mi}(\tau)$'s, and therefore the $\alpha_i(t_f)$'s, can be obtained by means of n simultaneous, uninterrupted backward integrations of the adjoint equations, where the mth integration is carried out for the mth set of initial conditions contained in the $\lambda_{mi}(t_f)$ tensor.

In the most general cases (where $\lambda_{mi}(t_f)$ represents a collection of n independent, but otherwise arbitrarily selected, sets of initial λ variables) Eqs. (A10) are completely coupled, i.e., every $\alpha_i(t_f)$ appears in each equation. However, a judicious choice for the tensor $\lambda_{mi}(t_f)$ removes the difficulty, for if:

$$\lambda_{mi}(t_f) = \delta_{mi} = \begin{cases} 0, & m \neq i \\ 1, & m = i \end{cases} \tag{A11}$$

then Eqs. (A10) simplify to:

$$\sum_{i=1}^{n} \lambda_{mi}(\tau)\alpha_i(\tau) = \alpha_m(t_f); \qquad m = 1, n \tag{A12}$$

which are completely uncoupled algebraic expressions for the $\alpha_m(t_f)$'s. Eq. (A12) is nothing more than the readily recognizable standard form for the solution of the set of linear homogeneous differential equations, (A1).

Acknowledgment

The author wishes to express his gratitude to Dr. Henry J. Kelley, Assistant Chief of Research, Research Department, Grumman Aircraft Engineering Corporation, whose suggestions formed the basis for a major portion of the substance of this text.

References

1. Battin, R. H.; "A Comparison of Fixed and Variable Arrival Navigation for Interplanetary Flight," Fifth Symposium on Ballistic Missile and Space Technology, at USC, Los Angeles, August 29–31, 1960.

2. Battin, R. H. and Laning, J. H., Jr.; "A Navigation Theory for Round-Trip Reconnaissance Missions to Venus and Mars," Fourth AFBMD-STL Symposium on Ballistic Missile and Space Technology, at UCLA Los Angeles, August 24–27, 1959.
3. Breakwell, J. V.; "Fuel Requirements for Crude Interplanetary Guidance," *Preprint 59-5* of the Western National Meeting of the A.A.S. in Los Angeles, August 1959.
4. Breakwell, J. V.; "The Spacing of Corrective Thrust in Interplanetary Navigation," *Preprint 60-76* of the Western National Meeting of the A.A.S. in Seattle, 1960.
5. Pfeiffer, C. G.; "Guidance for Space Missions," *External Publication No. 656*, Jet Propulsion Laboratory, Pasadena, California, June 23, 1959.
6. Moulton, F. R.; *Celestial Mechanics*, Second Edition, The Macmillan Company, New York, 1914.
7. Kelley, H. J.; "Gradient Theory of Optimal Flight Paths," *Preprint 1230-60* of the ARS Semi Annual Meeting, Los Angeles, California, May 9–12, 1960.
8. Kizner, W.; "Perturbation Theory and Green's Functions," *Section Report No. 12-152*, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, October 21, 1957.
9. Bliss, G. A.; *Mathematics for Exterior Ballistics*, John Wiley and Sons, Inc., New York, 1944.
10. Baker, R. M. L., Jr. and Makemson, M. W.; *An Introduction to Astrodynamics*, Academic Press, New York and London, 1960.
11. Goursat, E.; *A Course in Mathematical Analysis*, Part II, Volume II—Differential Equations, Ginn and Company, Boston, 1917.
12. Goodman, T. R. and Lance, G. N.; *The Numerical Integration of Two Point Boundary Value Problems*, Mathematical Tables and Other Aids to Computation, Vol. 10, No. 54, April 1956.

A Study of the Martian Upper Atmosphere and Ionosphere[1]

Gilbert Yanow[2]

Abstract

Hypothetical Martian models composed of 90–98% N_2 and 10–2% CO_2 with varying surface temperatures and densities have been selected and a probable photochemical chain determined. This photochemistry has been solved for incremental layers from the altitudes of 650 to 250 km using the most recent solar data extrapolated to the mean distance of Mars. The results are presented in graphical form for density, temperature, mean molecular mass, electron density, and atmospheric gas and ion concentrations against altitude as a function of two and three body recombination. The results at present are preliminary, and while these data are admittedly speculation, it appears plausible to suggest the existence of a fairly dense, multi-layered, high altitude ionosphere about Mars, further suggesting that terrestrial communication techniques could be employed using, however, perhaps different critical frequencies and skip distances.

Introduction

As the time when Man will actively explore the planets draws near, the need for upper atmosphere data of these bodies continually increases. Unfortunately, the amount of observed and experimental quantities of these data is limited. This situation has prompted the author to modify the astrophysical techniques used in stellar model calculations for determination of atmospheric models.

The method of approach is quite general in nature. A probable selection of possible chemical compositions for Mars has been taken from the findings of astronomical observers. After a careful analysis, a chain of likely photochemical reactions for the selection was determined. Using the latest solar data available, the solutions to the photochemistry were integrated down through the media from 650 to 250 km. The parameters of surface conditions, recombination process, and chemical composition were varied to determine how critically they effect the photochemical solutions.

It should be noted, that the numerical results are obviously quite speculative at the present time; they do, however, perhaps give an indication of which parameters are critical and some first order approximation to them.

Surface Conditions and Chemical Composition

During the past years many observers have made estimates of the conditions on Mars. Recently de Vaucouleurs [1] has made an extensive survey of the field and determined the best values. This information appears in Tables 1 and 2.

The surface pressure value can be derived by using the techniques of molecular light scattering. It is assumed that the atmospheric temperatures of the Earth and Mars are similar, and that the two atmospheres have the same approximate scattering properties for a certain spectral region ($\lambda > 0.5\ \mu$). The scattering coefficient ratio is then approximately equal to the gas mass ratio, and the surface pressure ratio is accordingly obtained by multiplying the mass ratio times the gravity ratio. The density value is directly obtained from these data.

[1] Submitted March 28, 1961.

[2] Missile and Space Systems, Douglas Aircraft Company Inc., Santa Monica, California.

TABLE 1

Environmental Conditions of Mars

surface pressure	$85 + 4$ mb $= 4 \times 10^3$ dyne/cm^3
surface density	$\doteqdot 10^{-4}$ gm/cm^3
surface temperature	$\doteqdot 273$°K mean (sub-solar point)
surface gravity	372.5 cm/sec^2

TABLE 2

Martian Atmospheric Chemical Compositions

Gas	Volume (%)
N_2	93.8
O_2	0.1
A	4.0 (?)
CO_2	$\doteqdot 2.2$

The surface temperature may be calculated by considering Mars as a gray sphere and applying Stefan's law using the solar constant. Estimates of the surface temperature have also been obtained from radiometric observations. The chemical composition, as given in Table 2, is based primarily on theoretical dynamics and chemistry. However, during observations of 1947–48 Kuiper identified CO_2 bands in the near-infrared region of the planet's spectrum.

Solar Data

Table 4 lists the majority of important solar lines in the region between 304 Å and 1817 Å. For the particular cases of interest in this study, additional information is given. The solution to the general problem is satisfied using the group of lines from 1808 through 1400 Å, the 1216 Å line, the pair centered at 618 Å, the 584 Å line, and the 304 Å line.

Photochemistry

As a basis for calculation, a nitrogen-carbon dioxide atmosphere has been assumed. A possible chain of photochemistry for such an atmosphere is shown in Table 3.

A rate coefficient of 10^{-6}cm^3/sec [2] was used for reaction (1). The cross-section for this reaction was taken as 3.7 ± 10^{-12} cm^2 [3]. Values of the rate coefficients for reactions (2a) and (6a) were not available. Consequently, the rather standard rate of 10^{-17} cm^3/sec was adopted. For the reactions (2b), (5), and (6b) the temperature dependent relation $5 \times 10^{-34}\ T^{1/2}$ cm^6/sec [4] was assumed as a rate coefficient. (It should be noted that the nitrogen atoms in reaction (6b) must be in an excited state; the corresponding reaction for atoms in the ground state is forbidden.) The nitrogen molecules are assumed to follow a dissociative recombination process as is illustrated by reactions (3) and (4). The cross-section used for (3) was 2×10^{-17} cm^2 [5] and the rate coefficient used for (4) was 10^{-6} cm^3/sec [6]. The cross-section and rate coefficient applied in reaction (7) were 5×10^{-18} cm^2 [7] and 10^{-7} cm^3/sec [6] respectively.

TABLE 3

Possible Photochemistry for a Nitrogen-Carbon Dioxide Atmosphere

(1) $CO_2 + h\nu \rightarrow CO + O$
(2a) $CO + O \rightarrow CO_2 + h\nu$
(2b) $CO + O + M \rightarrow CO_2 + M$
(3) $N_2 + h\nu \rightarrow N_2 + e$
(4) $N_2 + e \rightarrow N' + N''$
(5) $N + O + M \rightarrow NO + M$
(6a) $N' + N' \rightarrow N_2 + h\nu$
(6b) $N + N + M \rightarrow N_2 + M$
(7) $NO + h\nu \rightarrow N + O$
(8) $CO_2 + h\nu \leftrightarrow CO_2^+ + e$
(9) $O + h\nu \leftrightarrow O^+ + e$
(10) $N + h\nu \leftrightarrow N^+ + e$
(11) $NO + h\nu \leftrightarrow NO^+ + e$
(12) $CO + h\nu \leftrightarrow CO^+ + e$

Reactions (8) through (12) describe the photochemistry which determines the density and extent of the model Martian ionosphere. The cross-sections used in reactions (8) through (11) were, (8)3.3×10^{-17} cm^2 [8]; (9) 1.2×10^{-17} cm^2 [5]; (10) 10^{-17} cm^2; and (11) 2.2×10^{-18} cm^2 [10]. The recombination coefficients employed for the reactions were, (8) 10^{-7} cm^3/sec [2]; (9) 2×10^{-12} cm^3/sec [11]; and (11) 10^{-6} cm^3/sec

TABLE 4

Solar Emission Data [12] [13]

Line	Wave Length	Power Density Outside Earth	Power Density Outside Mars
	Å	erg/cm^2 sec	quanta/cm^2 sec
He II	304	1	6.58×10^9
He I	584	0.05	6.32×10^8
Mg X	610	0.05	7.55×10^8
Mg X	625	0.01	1.51×10^8
H Lyman ϵ	938	0.004	—
H Lyman δ	950	0.002–0.003	—
H Lyman γ	973	missing*	—
H Lyman β	1026	0.06	—
Si III	1207	0.2	—
H Lymanα	1216	6	—
Si II	1265	0.02	—
O I	1302	0.04	—
O I	1305	0.04	—
O I	1306	0.04	—
Si II	1309	0.04	—
C II	1335	0.3	—
C II	1336	0.3	—
Si IV	1394	0.2	—
Si IV	1403	0.2	
Si II	1527	0.03	
Si II	1533	0.06	
C IV	1548	0.6	
C IV	1551	0.3	
C I	1560	0.12	5×10^{11}
He II	1640	0.3	
C I	1656	0.3	
C I	1658	0.2	
Fe II	1670	0.2	
Si II	1808	1.1	
Si II	1817	1.7	

* It is believed that H Lyman γ is absorbed by the Earth's atmosphere.

[10]. The cross-section of reaction (12) was assumed to be the same as reaction (8), and the recombination coefficients of reaction (12) and (11) were assumed the same.

Reaction (8) took precedence over reaction (1) when the wave length was equal to or less than 885 Å. The same condition held for reaction (11) over reaction (7) when the wave length equaled or was less than 1348 Å.

At present, values of rate coefficients, cross-sections, and recombination coefficients are not readily available. Consequently, some of the data used in the calculations were not taken directly from the quoted reference, but may have been either an extrapolation or a value assumed as characteristic of the particular reaction. No information was found for the spectral region of approximately 500 Å or less. It is known, however, that as the wave length approaches zero, the cross-section also goes to zero, and it appears justifiable to assume a drop in the value of the cross-section of about one order-of-magnitude at the spectral region of the HeII line, 304 Å. This assumption was made.

Method of Solution

Assume that it were possible to insulate the model planet Mars from solar radiation. After a sufficient length of time, all the electrons, ions, and atoms in the atmosphere would recombine to form only molecular nitrogen and carbon dioxide. Since for all practical purposes the ionosphere photochemistry may be considered reversible, this equilibrium condition may be regarded as the state from which the ionosphere was formed.

Initial Density Distribution

It is assumed that the "insulated" atmosphere behaves isothermally, is in hydrostatic equilibrium, and the acceleration of gravity follows an inverse square relation; the density may be given by the customary density equation,

$$\rho = \rho_0 \exp -(mg_0/kT)(ah/a + h), \qquad (1)$$

where

ρ_0 = surface density,

m = mean molecular mass of the gas,

g_0 = surface acceleration of gravity,

k = Boltzmann constant,

T = abolute temperature,

a = radius of the planet, and

h = altitude above the surface.

It is further considered that in the upper regions of the model atmosphere the constituent gases will not be well mixed because of a low amount of convection, but will take on a distribution that is a function of their respective mean molecular mass, the gravity gradient, and the kinetic temperature. This distribution is approximated by using Eq. (1) with the surface density, ρ_0, equal to the fraction of the constituent gas time the overall density. For example, if the gas were CO_2 and it amounted to 5% of the total atmosphere, the ρ_0 used would equal 0.05×10^{-4} gm/cm^3 for an overall surface density of 10^{-4} gm/cm^3.

Photochemical Analysis

The number of photochemical reactions of any particular type may be expressed

$$q_{hi} = A_i N_{hi} Q_\lambda , \qquad (2)$$

where

q_{hi} = number of photochemical process occurring of specied i at altitude h,

A_i = cross-section of process of species i,

N_{hi} = neutral particle density of species i at altitude h, and

Q_λ = number of quanta of wavelength λ available per sec.

Using Beer's Law, for normal incidence

$$\Delta Q_\lambda = A_i N_{hi} Q_\lambda \Delta h \qquad (3)$$

and

$$Q_\lambda = Q_{\lambda_0} \exp \sum_h \sum_i - A_i N_{hi} \Delta h, \qquad (4)$$

where $Q_{\lambda_0} = Q_\lambda$ at some datum level.

A combination of Eqs. (4) and (1) yields

$$q_{hi} = A_i N_{hi} Q_{\lambda_0} \exp \sum_h \sum_i - A_i N_{hi} \Delta h. \qquad (5)$$

Equation (5) can be applied directly to reactions (1) through (7). In reactions (8) through (12) electron-ion recombination is also a consideration. Bates and Massey [11] have shown that

$$dN/dt = (q/1 + \beta) - (\alpha_e + \beta\alpha_i)N_e^2 - (N_e/1 + \lambda)\, d\beta/dt, \qquad (6)$$

where

dN/dt = rate of change of the neutral particle, density or the rate of production of electrons,

N_e = electron density, and

α_e, α_i, β = coefficients of various modes of recombination.

Taking steady-state conditions, i.e.,

$$dN/dt = d/dt = 0,$$

Eq. (6) reduces to

$$q/1 + \beta = \alpha N_e^2, \qquad (7)$$

where α = effective recombination coefficient. As a general rule $\beta \ll 1$, so Eq. (7) may be written

$$N_e = (q/\alpha)^{\frac{1}{2}}. \qquad (8)$$

Equation (8) may be applied to reactions (8) through (12). The photochemical differential equation of reactions (1) and (2a) to determine the amount of CO and O at altitude h is given by

$$d(\mathrm{CO})/dt = d(\mathrm{O})/dt = K_1 I_1 - K_{2a} C_{CO} C_O = 0, \tag{9}$$

where

K_1, K_{2a} = rate coefficients,
I_1 = photons or quanta absorbed per sec ($I = q_{CO_2}$), and
C_{CO}, C_O = concentrations of CO and O.

The solution to this equation is

$$C_{CO} = C_O = [K_1/K_{2a}(N_{CO_2} A_\lambda Q_\lambda)]^{1/2}. \tag{10}$$

In a similar manner the solution for reactions (3), (4), (5), and (6a) is

$$C_N = \{-K_5 C_O C_M + [4K_{6a}K_4(C_{N_2} A_\lambda Q_\lambda)^{1/2}]^{1/2}\}/2K_{6a}, \tag{11}$$

where C_m = third body concentration. The solutions to reactions (8) through (12) all have the form of Eq. (8).

Boundary Conditions

Mitra [14] has shown that an approximate value for the critical density or the density as a function of altitude where the atmospheric gases will just start to escape, is given by

$$n_c \doteq [(mg_0 r_0/kT) - 2]/\pi\sigma^2 r_0, \tag{12}$$

where

σ = molecular diameter, and
r_0 = distance from the center of the planet to the point, where the atmosphere begins to behave isothermally.

Using the values of
$m = 28 \times 1.67 \times 10^{-24}$ gm
$g_0 = 372.5$ cm/sec^2
$k = 1.38 \times 10^{-16}$ erg/°K
$T = 1500$°K
$r_0 = 3410$ km (3400 km radius plus 10 km tropopause)
$\sigma = 5 \times 10^{-8}$ cm (after Mitra, 1952).

The value of n_c, 4.7×10^{-16} gm/cm^3, obtained at the approximate altitude of 640 km, was adopted. Below the approximate region of 250 km recombination increases rapidly so that for all practical purposes the ionosphere has ceased to exist.

Calculations

The results presented in this paper have been calculated using a Bendix G-15 digital computer. The technique used is approximately the following. An initial density at a particular incremental layer, for example 645 km to 650 km, was found using Eq. (1). Then using the energy of the 1216 Å region, reactions (1) and (2a) were solved for the CO, O, and CO_2 equilibrium condition concentrations. By successive use of the 618 Å, 584 Å, and 304 Å energies, reactions (3), (4), (5), and (6a) (or perhaps reaction, (3), (4), (5), and (6b)) were solved for the N, N_2, O, and NO concentrations. The 1808-1400 Å, 618 Å, 584 Å, and 304 Å energies were then used in turn to solve for the ion-neutral equilibrium balance in reactions (8) through (12). When a photoreaction took place it was considered that energy was used to dissociate, ionize, and in certain instances excite the atoms produced. After these sequences occurred there was a certain amount of energy in excess of what was used for each event. For example, using the 618 Å energy with reactions (3) and (4) there was 0.7 eV extra per event, with 584 Å 1.7 eV, and with 304 Å 17.8 eV. This excess energy must go into a kinetic form, and after a sufficiently long period of time, this kinetic energy should be fairly evenly distributed over the particles of the particular incremental layer. A simple "$\frac{3}{2}kT$" relationship was then used to calculated the increase in the temperature over the isothermal value produced by the photochemistry. The gas in the incremental layer was assumed to behave as a perfect gas and further that the temperature was the only parameter to change. Consequently, the effect that the temperature change had upon the density was to decrease the density by the ratio of the original temperature over new temperature. A new density was obtained in this manner, and the whole calculation of the photochemistry was repeated.

Equation (4) was then used to determine the transmitted energy to the next lower layer, and using this energy the above procedure was carried through for this lower increment. In this manner the problem was numerically integrated from 650 km to 250 km.

Results and Comments

Figure (1) is a density vs altitude curve as a function of recombination process. The curve is based on a surface density and temperature of 10^{-4} gm/cm^3 and 273°K respectively. From about 350 km to the lower

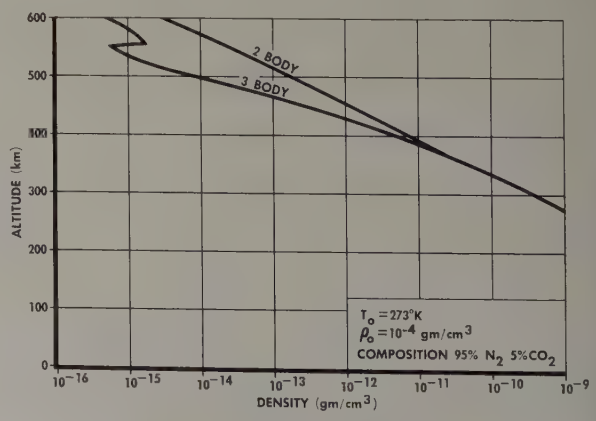

FIG. 1

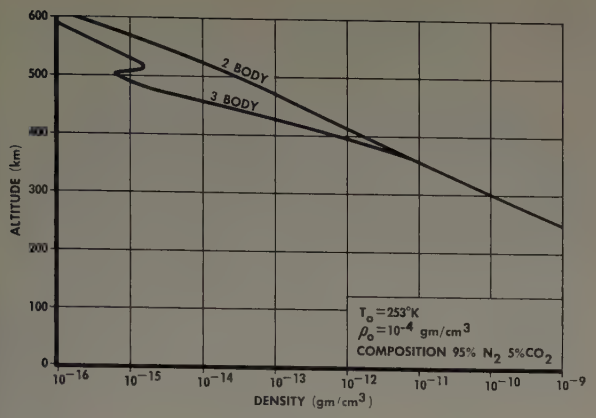

FIG. 2

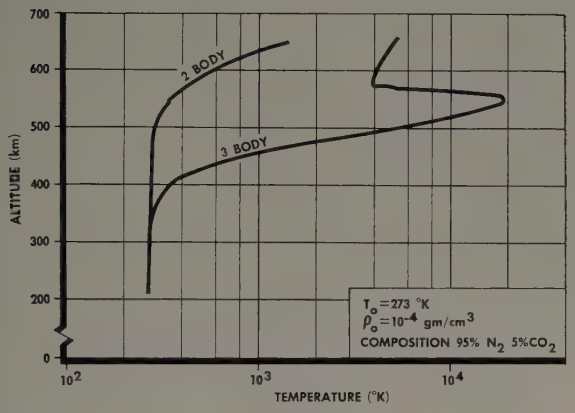

FIG. 3

limit, the density value appears the same regardless of whether two body or three body recombination is used. However, in the upper regions the two body process produces a somewhat greater density. Also it is noticed that the two body curve does not display the discontinuity shown by the three body curve at the altitude of about 500 km.

Figure (2) is similar to Fig. (1) with the only exception being that the calculations were based on a surface temperature of 253°K rather than 273°K. The shapes of the curves of the two figures are the same. The only difference the twenty degree drop in the surface temperature makes is to lower the curve, point for point, about 50 km. When the surface temperature is raised, the curves are displaced upward by a similar amount. A change in the surface density has the same effect upon the solution as a change in the surface temperature.

Figure (3) is a relation between temperature and altitude using the surface temperature and density of 273°K and 10^{-4} gm/cm^3 respectively. Again the three body curve exhibits a rather large discontinuity at about 500 km, indicating temperatures that are perhaps unrealistic. The two body curve, on the other hand, represents values which are considerably more in the realm of reality.

Figure (4) is also a temperature vs altitude graph but based on surface conditions of 253°K and 10^{-4} gm/cm^3. It shows the same characteristics as described in relation to Fig. (2).

A curve of mean molecular mass against altitude is displayed in Fig. (5). From approximately 400 km downward both recombination methods appear to produce an atmosphere composed of primarily molecular nitrogen. Above this altitude, however, the composition is strongly dependent on whether two body or three body recombination is employed. Using the latter, the prime constituent is atomic nitrogen chile with the former the model atmosphere continues to be composed of primarily molecular nitrogen.

Figure (6) represents a profile of the model Martial ionosphere. Only data based on a chemical composi-

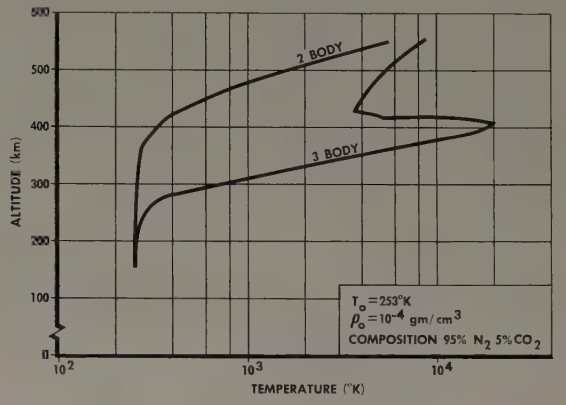

FIG. 4

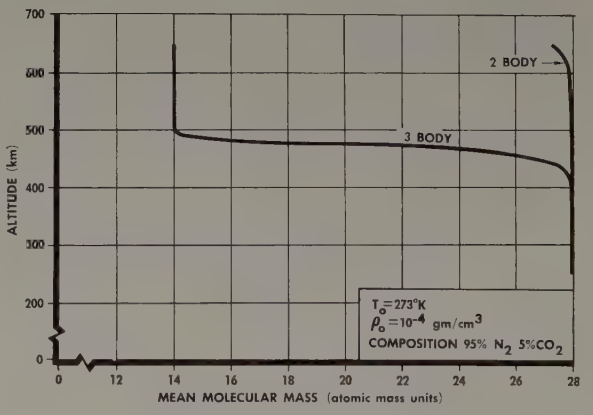

FIG. 5

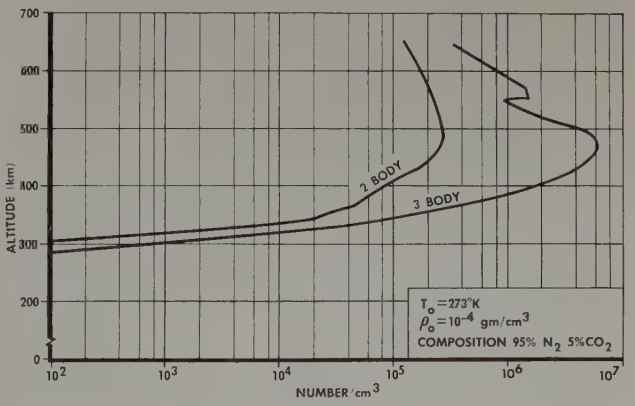

FIG. 6

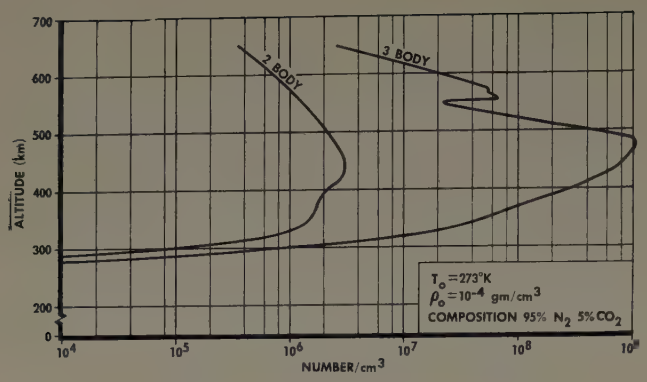

FIG. 7

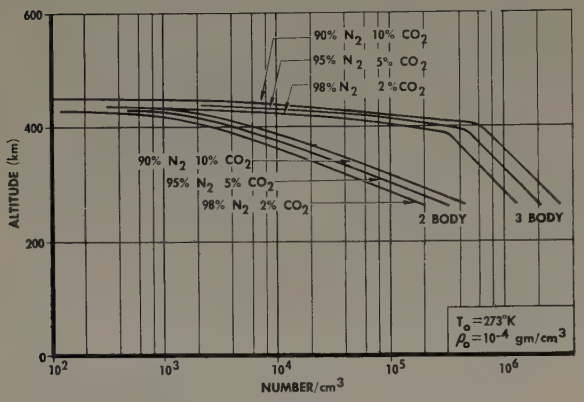

FIG. 8

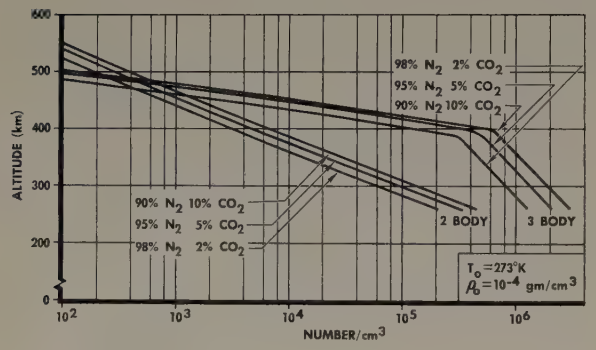

FIG. 9

tion of 95 % N_2 and 5 % CO_2 are shown. The variation of chemistry within the limits used had very little effect upon the solutions. The concentrations of N^+ are almost identical to these curves, the only small departure occurring with the two body curve from about 400 km downward. As is demonstrated, the choice of recombination process strongly effects the numerical results.

The three body process produces an ionosphere of two layers, separated by a marked discontinuity. From an analogy with the Earth, the uppermost peak was designated the F_2 layer and the lower peak the F_1 layer. The model F_2 and F_1 layers occur at approximate altitudes of 560 km and 470 km respectively, with corresponding densities of approximately 10 electron/cm^3 and 6×10^6 electron/cm^3. For comparison, the Earth's F_2 and F_1 layers are found at the alti-

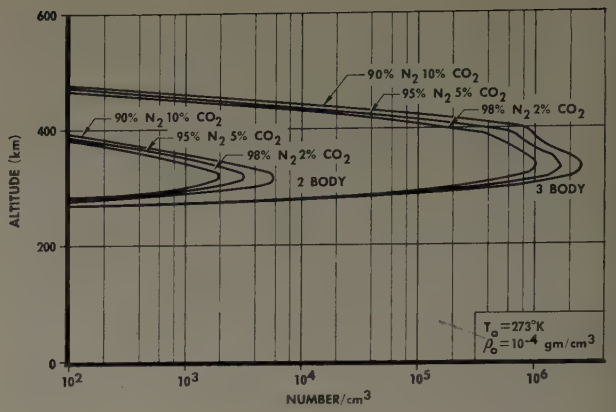

FIG. 10

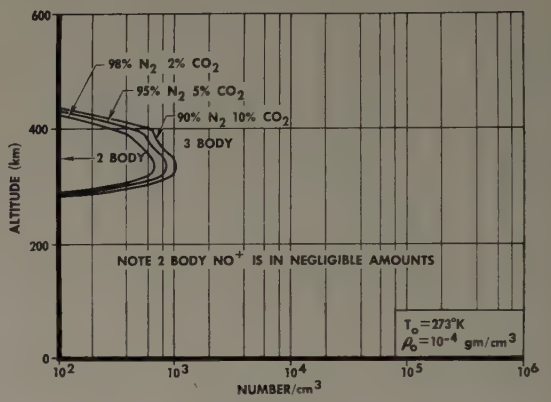

FIG. 11

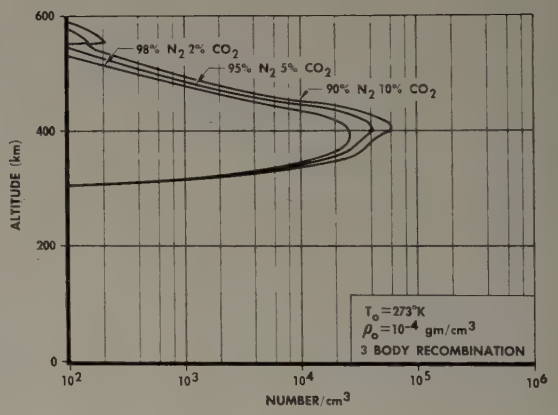

FIG. 12

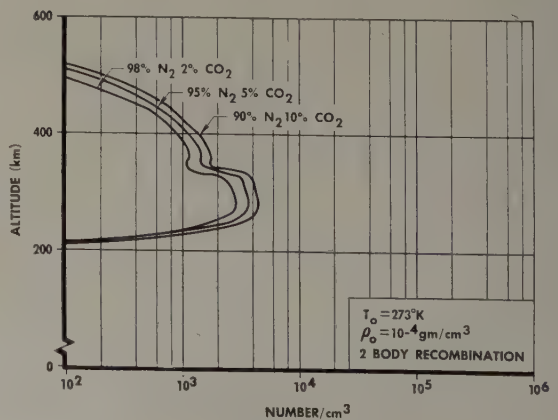

FIG. 13

tudes of about 300 km and 200 km with respective densities of the order of 10^6 electron/cm^3 and 10^5 electron/cm^3.

The two body model calculations also produce a two layer ionosphere. However, the layers are not nearly so obviously separate. These F_2 and F_1 peaks are at the elevations of about 430 km and 360 km, with density values of 3×10^5 electron/cm^3 and 4×10^4 electron/cm^3 approximately. It should perhaps be noted that even when a recombination coefficient as large as 10^{-16} cm^3/sec and a surface density of 10^{-5} gm/cm^3 were used, peaks of the order of 10^5 electron/cm^3 and 10^4 electron/cm^3 were still obtained for the calculated F_2 and F_1 regions.

Figure (7) is a curve of atomic nitrogen concentration with altitude as a function of the recombination process. The results of this graph seem to indicate that in the regions of the atmosphere above 300 km, for the surface conditions used, two body recombination is far more efficient than is three body recombination.

Figures (8) and (9) are respective graphs of the atomic oxygen and carbon monoxide concentrations against altitude, depending on the recombination process employed. Of particular interest is the fact that the curves appear to indicate that the form of recombination is not of cardinal importance near the regions of the upper or lower limits of integration but is of importance in the mid regions of about 400 km altitude.

Nitrogen oxide vs altitude as a function of recombination process is shown in Fig. (10). With a two body process the NO concentration is far less than with three body recombination. The two body NO layer is also much thinner than the three body layer, but both concentration curves peak at about the same altitude.

Figure (11) is a curve of the nitrogen oxide ion concentration. Only a three body recombination mode produces any reasonable amounts of NO^+.

The concentrations of the oxygen ion, depending on whether three body or two body recombination was used, are presented in Figs. (12) and (13) respectively. The two body process yields less O^+ than does the three body, and the peak of the two body curve occurs at an altitude of about 100 km lower than the corresponding three body curve. Further, while the two body process appeared to narrow the NO^+ concentration, the opposite is the case with the accounts of 0^+.

Conclusions

Regardless of which recombination process was used, the two body or the three body, the results consistently indicated the possible existence of a multilayered, high altitude, fairly dense ionosphere about the planet Mars, if the atmosphere is composed of primarily nitrogen. The author feels that the results of the work show this technique of study promising, and is at present attempting to start the task of model Earth calculations. A better evaluation of the procedure can then be made with a comparison of the Earth study results with known terrestrial data.

Acknowledgment

The author gratefully acknowledges the encouragement given him during the preparation of this paper by Dr. Gerard de Vaucouleurs and for his many helpful criticisms.

References

[1] Vaucouleurs, G. de, *Proceeding of the 2nd Annual Symposium of Physics and Medicine of Space*, Antonio, Texas, November 1958.

[2] Bates, D. R. and Witherspoon, A. E., "The Photo-Chemistry of Some Minor Constituents of the Earth's Atmosphere (CO_2, CO, CH_4, N_2O)," *Roy. Astr. Soc. Monthly Notices*, **112,** 101 (1952).

[3] Wilkinson, P. G. and Johnston, H. L., "The Absorption Spectra of Methane, Carbon Dioxide, Water Vapor, and Ethylene in the Vacuum Ultraviolet," *Journ. Chem. Phy.*, **18,** 2, 190 (1950).

[4] Bates, D. R., "The Physics of the Upper Atmosphere," *The Earth as a Planet*, page 576, edited by G. P. Kuiper, Univ. Chicago Press, 1954.

[5] Bates, D. R. and Seaton, M. J., "The Quantal Theory of Continuous Absorption of Radiation by Various Atoms in their Ground States. II Futher Calculations on Oxygen, Nitrogen, and Carbon," *Roy. Astr. Soc. Monthly Notices*, **109,** 6, 698, 1948.

[6] Nicolet, M., "Dynamic Effects in the Upper Atmosphere," *The Earth as a Planet*, page 644, edited by G. P. Kuiper, Univ. Chicago Press, 1954.

[7] Watanabe, K., Marmo, F. F., and Inn, E. C. Y., "Photoionization Cross Section of Nitric Oxide," *Phy. Rev.*, **91,** 5, 1155, 1953.

[8] Wainfan, N., Walker, W. C., and Weissler, G. L., "Photoionization Efficiencies and Cross Sections in O_2, N_2, CO_2, A, H_2O, and CH_4," *Phy. Rev.*, **99,** 2, 542, 1955.

[9] Ehler, A. W. and Weissler, G. L., "Ultraviolet Absorption of Atomic Nitrogen in its Ionization Continum," *Jour. Opt. Soc.*, **45,** 12, 1035, 1955.

[10] Marmo, F. F., Pressman, J., Aschenbrand, L. M., Jursa, A. S., and Zelikoff, M., "The Formation of an Artificial Ion Cloud; Photoionization of NO by Solar Lyman Alpha at 95km," *The Threshold of Space*, page 232, edited by M. Zelikoff, Pergamon Press, 1957.

[11] Bates, D. R. and Massey, H. S. W., "Basic Reactions in the Upper Atmosphere I," *Proc. Roy. Soc.*, **187,** 261, 1946.

[12] Aboud, A., Behring, W. E., and Rense, W. A., "Emission Intensities in the Solar Ultraviolet," *Astrophy. Journ.*, **130,** 2, 381, 1959.

[13] Naval Research Laboratory, personal communication.

[14] Mitra, S. K., *The Upper Atmosphere*, Asiatic Society, 1952.

[15] Kallmann, H. K., *A Study of the Ionosphere*, Ph.D. Thesis, U.C.L.A., 1955.

[16] Herzberg, G., *Atomic Spectra and Atomic Structure*, Dover, 1944.

[17] Herzberg, G., *Spectra of Diatomic Molecules*, Van Nostrand, 1957.

[18] Hess, S. L., "Some Aspects of the Meteorology of Mars," *Journ. of Meteorology*, **7,** 1, 1950.

[19] Goody, R. M., "The Atmosphere of Mars," *Journ. of Brit. Interplan. Soc.*, **16,** 2, 17, 1957.

Attitude Stability of an Elastic Body of Revolution in Space[1]

Leonard Meirovitch[2]

Abstract

The motion of a body in space is described by three angular velocities called spin, precession, and nutation.

A rigid body of revolution with no external moments has an angular momentum vector, **h**, constant in space and its spin or symmetry axis precesses about the vector **h** at a constant attitude angle θ. An elastic body, however, dissipates energy and as a result the angle θ changes. If the decrease of angle θ denotes stability, it is shown on basis of energy considerations that stability is implied by $C/A > 1$ and instability by $C/A < 1$, where C and A are mass moments of inertia about the spin and pitch axes respectively.

A body consisting of two elastic disks connected by a rigid shaft so that $C/A < 1$ is considered. Using the elastic solution of a disk subjected to gyroscopic forces the time t required to reach an attitude angle θ is obtained.

The results are compared with other investigations of the same subject and the agreement is found very satisfactory.

SYMBOLS

$\mathbf{h}$	= angular momentum vector
θ	= attitude angle measured from vector **h** to spin axis
A, C	= moment of inertia about pitch and spin axes respectively
U	= strain energy
γ_E	= hysteretic damping factor for normal stresses
σ	= stress
E	= Young's modulus
V	= volume
T	= kinetic energy
ω_0	= initial angular velocity
t	= time
$\dot{\theta}$	= nutation velocity
D_E	= modified plate flexural rigidity
r, α	= polar coordinates
ρ	= mass density
$\dot{\phi}$	= spin velocity
D	= plate flexural rigidity
w	= plate deflection
ν	= Poisson's ratio
h	= plate thickness
$W(r)$	= amplitude of plate deflection
γ_G	= hysteretic damping factor for shearing stresses
t_0	= cyclic time
a, b	= external and internal radii of the disk

Introduction

The motion of a moment-free rigid body with principal mass moments of inertia A, B, C is an unsteady periodic precession and nutation about the resultant angular momentum vector **h** which is fixed in space. Steady motion is possible only about the axes of maximum or minimum moment of inertia. A rigid body of revolution A, A, C however has no intermediate moment of inertia and, therefore, the moment-free motion consists of steady precession of the axis of symmetry about the fixed angular momentum vector **h**. The angle θ between the symmetry or spin axis and the direction of the **h** vector is called attitude angle and is a constant [1].

A real body, however, is elastic. A vibrating elastic body undergoes stresses and deformations resulting in energy dissipation which in turn causes a change in the attitude angle θ. The forces causing the elastic vibrations are the gyroscopic or inertia forces. The rate of energy dissipation per cycle of stress is the area within the hysteresis loop as shown in Fig. 1.

It is customary to assume that the energy dissipated in one cycle is a fraction of the strain energy. This is obtained by multiplying the strain energy expression by a coefficient γ_E which is called the hysteretic damping factor.

In general, for a body subjected to stress of one kind only, namely σ, the energy dissipation per cycle is given by

$$U_{\text{dis}} = \frac{1}{2}\gamma_E \iiint_{\text{volume}} \frac{\sigma^2}{E}\, dV. \tag{1}$$

The change in the attitude angle θ due to energy dissipation can be of a stabilizing nature or the opposite. It follows that for an elastic body where energy is dissipated during vibration, the concept of stability of motion should be revised. It was shown [1], that the expression of the kinetic energy rate of change is

$$\frac{dT}{dt} = C\omega_0^2 \left(\frac{C}{A} - 1\right) \sin\theta \cos\theta \frac{d\theta}{dt}, \tag{2}$$

where

A = mass moment of inertia about the pitch axis,
C = mass moment of inertia about the spin axis, and
ω_0 = initial angular velocity.

Since energy is dissipated rather than gained, the kinetic energy must decrease. Hence, the right hand

[1] From a dissertation submitted in partial fulfillment for the degree Ph.D. in Engineering at the University of California in Los Angeles. Manuscript submitted August 30, 1961.

[2] Staff Engineer, IBM General Products Division, Development Laboratory, Endicott, N. Y.

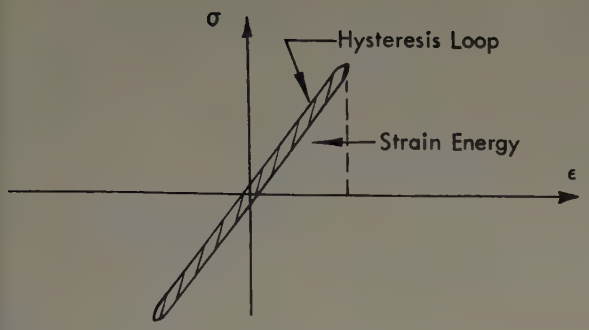

FIG. 1. Stress-strain diagram showing the hysteresis loop

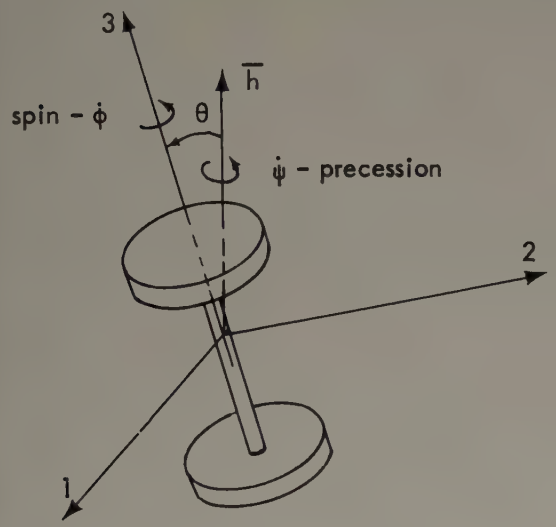

FIG. 2. The body and the angular velocities

side must be negative. There are two possible cases:

$$\text{(a)} \qquad \frac{C}{A} - 1 > 0 \quad \text{and} \quad \frac{d\theta}{dt} < 0$$

and (3)

$$\text{(b)} \qquad \frac{C}{A} - 1 < 0 \quad \text{and} \quad \frac{d\theta}{dt} > 0.$$

From Fig. 2, one observes that instability is characterized by a continuous increase of the initially small angle θ.

Hence,

$C > A$ implies stability
$C < A$ implies instability.

The above principle has been used in devices which stabilize satellites of the first type by intentional dissipation of energy.

The problem to be discussed here treats the case shown in Fig. 2, where the two circular disks are elastic and the connecting shaft is assumed rigid. The elastic solution of the elastic disk subjected to gyroscopic forces was described in a previous paper [2]. Using the expression of the deflection, w, derived in the above paper, an expression for the strain energy is obtained. By equating the rate of energy dissipation, $\dot{U}_{\text{dis}}$, to the rate of change, $\dot{T}$, of the kinetic energy, a relation between the rate of change, $\dot{\theta}$, of the attitude angle, θ, and the angle, θ, is obtained. Integrating the rate of change, $\dot{\theta}$, the time, t, to produce an attitude angle, θ, is determined.

The Kinetic and Strain Energy Expressions

In the first reference, it was shown that the kinetic energy expression of a body of revolution is

$$T = \frac{1}{2} C\omega_0^2 \left[1 + \left(\frac{C}{A} - 1\right) \sin^2 \theta\right]. \tag{4}$$

The second reference shows that the deflection of the circular elastic disk subjected to gyroscopic forces is given in polar coordinates by

$$w = \Big[X(1,1)J_1(\gamma r) + X(2,1)Y_1(\gamma r) + X(3,1)I_1(\gamma r) + X(4,1)K_1(\gamma r) + \frac{f}{\sqrt[4]{\gamma_{D_E}}}\, r\Big] \cdot \sin(\alpha + \dot{\phi}t). \tag{5}$$

In the above $J_1(\gamma r)$ and $Y_1(\gamma r)$ are Bessel functions of the first order and first and second kind respectively. $I_1(\gamma r)$ and $K_1(\gamma r)$ are modified or hyperbolic Bessel functions of the first and second kind respectively. The coefficients $X(j, 1)$ were determined from the boundary conditions. In addition

$$\gamma^4 = \frac{\rho}{D_E} \dot{\phi}^2 \tag{6}$$

and

$$f = \rho\omega_0^2 \frac{C}{A}\left(2 - \frac{C}{A}\right) \sin\theta \cos\theta. \tag{7}$$

The strain energy due to normal stresses is expressed in polar coordinates by

$$U_E = \frac{1}{2} D \iint_{\text{Area}} \left[(\nabla^2 w)^2 - 2(1-\nu)\frac{\partial^2 w}{\partial r^2} \cdot \left(\frac{1}{r}\frac{\partial w}{\partial r} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \alpha^2}\right)\right] r\, dr\, d\alpha, \tag{8}$$

where $D = hD_E = \dfrac{Eh^3}{12(1-\nu^2)}$ is the plate flexural rigidity. Since every point at an angle α goes through all values ranging from $-W(r)$ to $W(r)$ twice in one revolution, it follows that the strain energy per cycle of stress is obtained by writing expression (8) in terms of the amplitude, $W(r)$, of the deflection, w, and letting α vary between zero and 2π. Consequently the strain

energy due to normal stresses per cycle of stress is

$$U_{E/\text{cycle}} = \pi D \int_b^a \Big\{ \gamma^4 [X(1,1)J_1(\gamma r) + X(2,1)Y_1(\gamma r) - X(3,1)I_1(\gamma r) - X(4,1)K_1(\gamma r)]^2 r - 2(1-\nu)\gamma^3 \cdot \Big[X(1,1) \Big\{ -\frac{1}{\gamma r} J_0(\gamma r) + \Big[\frac{2}{(\gamma r)^2} - 1 \Big] \cdot J_1(\gamma r) \Big\} + X(2,1) \Big\{ -\frac{1}{\gamma r} Y_0(\gamma r) + \Big[\frac{2}{(\gamma r)^2} - 1 \Big] Y_1(\gamma r) \Big\} + X(3,1) \Big\{ -\frac{1}{\gamma r} I_0(\gamma r) + \Big[\frac{2}{(\gamma r)^2} + 1 \Big] I_1(\gamma r) \Big\} + X(4,1) \cdot \Big\{ \frac{1}{\gamma r} K_0(\gamma r) + \Big[\frac{2}{(\gamma r)^2} + 1 \Big] K_1(\gamma r) \Big\} \Big] \cdot \Big[X(1,1) \Big\{ J_0(\gamma r) - \frac{2}{\gamma r} J_1(\gamma r) \Big\} + X(2,1) \Big\{ Y_0(\gamma r) - \frac{2}{\gamma r} Y_1(\gamma r) \Big\} + X(3,1) \Big\{ I_0(\gamma r) - \frac{2}{\gamma r} I_1(\gamma r) \Big\} - X(4,1) \Big\{ K_0(\gamma r) + \frac{2}{\gamma r} K_1(\gamma r) \Big\} \Big] \Big\} dr. \tag{9}$$

In a similar way the expression for the strain energy due to shearing stresses per cycle of stress is written

$$U_{G/\text{cycle}} = 2\pi(1-\nu)D \int_b^a \frac{\gamma^2}{r} \cdot \Big\{ X(1,1) \Big[J_0(\gamma r) - \frac{2}{\gamma r} J_1(\gamma r) \Big] + X(2,1) \Big[Y_0(\gamma r) - \frac{2}{\gamma r} Y_1(\gamma r) \Big] + X(3,1) \Big[I_0(\gamma r) - \frac{2}{\gamma r} I_1(\gamma r) \Big] - X(4,1) \Big[K_0(\gamma r) + \frac{2}{\gamma r} K_1(\gamma r) \Big] \Big\}^2 dr. \tag{10}$$

Kinetic and Strain Energy Rates of Change

The rate of change of the kinetic energy is obtained by simply differentiating Eq. (4) with respect to time

$$\dot{T} = C\omega_0^2 \left(\frac{C}{A} - 1 \right) \sin\theta \cos\theta \dot{\theta}. \tag{11}$$

As mentioned in the first section, the energy dissipated in one cycle of stress is assumed to be a fraction of the strain energy. Specifically,

$$\Delta U_{/\text{cycle}} = \gamma_E U_{E/\text{cycle}} + \gamma_G U_{G/\text{cycle}}, \tag{12}$$

where $U_{E/\text{cycle}}$ and $U_{G/\text{cycle}}$ are given by Eqs. (9) and (10), and γ_E and γ_G are hysteretic damping factors for normal and shearing stresses respectively. To find the rate of change one has to divide expression (12) by the cyclic time t_0 which is the time required to perform one revolution. It was shown previously [1] that the spin velocity expression is

$$\dot{\phi} = \left(1 - \frac{C}{A} \right) \omega_0 \cos\theta, \tag{13}$$

so that the cyclic time is

$$t_0 = \frac{2\pi}{\dot{\phi}} = \frac{2\pi}{\left(1 - \frac{C}{A} \right) \omega_0 \cos\theta}. \tag{14}$$

It follows that the strain energy rate of change is

$$\dot{U} = \frac{\left(1 - \frac{C}{A} \right) \omega_0 \cos\theta}{2\pi} (\gamma_E U_{E/\text{cycle}} + \gamma_G U_{G/\text{cycle}}). \tag{15}$$

Attitude Angle as a Function of Time

The rate of change of the attitude angle θ is obtained by equating the rate of change of the kinetic energy to the rate of energy dissipation. In order to account for the sign one writes

$$\dot{T} = -\dot{U}, \tag{16}$$

and using expressions (11) and (15), Eq. (16) yields rate of change of θ

$$\dot{\theta} = \frac{1}{2\pi \omega_0 C \sin\theta} (\gamma_E U_{E/\text{cycle}} + \gamma_G U_{G/\text{cycle}}). \tag{17}$$

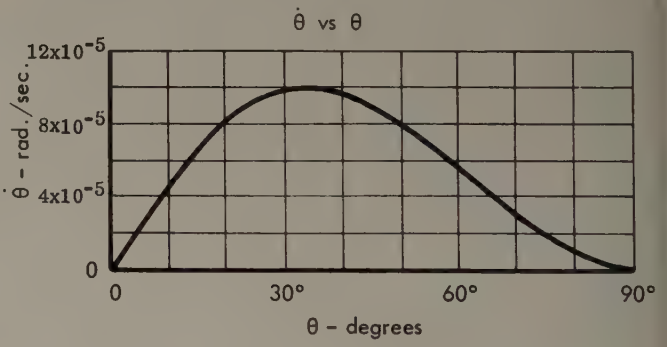

FIG. 3. Rate of change $\dot{\theta}$ as function of attitude angle θ

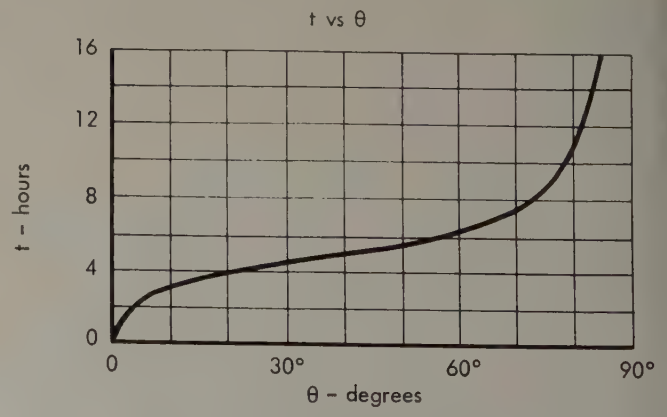

FIG. 4. Time t as function of attitude angle θ

Note that $U_{E/\text{cycle}}$ and $U_{G/\text{cycle}}$ are depending on θ. Equation (17) integrated gives the time t as a function of the attitude angle θ. Separating variables one obtains

$$t = 2\pi\,\omega_0\,C \int_0^{\theta} \frac{\sin\theta}{(\gamma_E\,U_{E/\text{cycle}} + \gamma_G\,U_{G/\text{cycle}})}\,d\theta, \quad (18)$$

where $U_{E/\text{cycle}}$ and $U_{G/\text{cycle}}$ are given by Eqs. (9) and (10).

Numerical Example

Using the configuration shown in Fig. 2 a numerical example was worked out. Data used is as follows:

Internal radius	b	= 4.0 in
External radius	a	= 20.0 in
Thickness	h	= 0.5 in
Mass density	ρ	= 7.3237×10^{-4} lb in^{-4} sec^2
Poisson's ratio	ν	= 0.30
Young's modulus	E	= 30×10^6 lb in^{-2}
Hysteretic damping factor for normal stresses	γ_E	= 0.1
Hysteretic damping factor for shearing stresses	γ_G	= 0.1
Initial angular velocity	ω_0	= 20 rad sec^{-1}
Moments of inertia ratio	C/A	= 0.1

Equation (17) in conjunction with expressions (9) and (10) were used to determine the rate of change $\dot{\theta}$ as a function of the attitude angle θ. The integrals in expressions (9) and (10) were evaluated numerically. The results are plotted in Fig. 3. In order to obtain the attitude angle, θ, as a function of the time, t, a procedure more suitable for computer use was adopted instead of equation (18). The function θ vs. t is exhibited in Fig. 4.

Conclusions

The results reported in the previous section indicate that the analysis presented is perfectly justified, and that the energy dissipation through cyclic stress is a factor which can become quite serious.

In order to obtain the solution of the plate the customary assumptions of elastic behavior and small deformations were made. Both were justified by the numerical results. In other cases, however, the assumptions may not hold and large deflections and stresses beyond the elastic limit may occur. Though an analytic solution under such conditions does not seem possible, the fact remains that such a problem may exist. Hence, for cases $C/A < 1$ a careful analysis has to be made.

The assumption that the nutation velocity, $\dot{\theta}$, is small compared to the spin velocity, $\dot{\phi}$, and the precession velocity, $\dot{\psi}$, appeared to be fully justified.

In the presently considered case, one can deduce from Fig. 4 that in a matter of hours the axis of symmetry deviates from the initial direction along the angular momentum $\mathbf{h}$ to an appreciable angle θ.

In general, the numerical results from the previous section conform quite well with the expectations. They also indicate excellent agreement with previous work done along the same line by Thomson as far as the dynamical behavior of the system is concerned.

Acknowledgment

The author wishes to thank Professor W. T. Thomson from the University of California, Los Angeles, both for suggesting the problem and for his advice during the course of the investigation.

References

[1] Thomson, W. T., and Reiter, G. S., "Attitude Drift of Space Vehicles," *The Journal of the Astronautical Sciences*, **VII**, Number 2, Summer, 1960.

[2] Meirovitch, L., "Bending Vibrations of a Disk Subjected to Gyroscopic Forces," *The Journal of the Astronautical Sciences*, **VIII**, Number 3, Fall, 1961.

[3] Morse, P. M., *Vibration and Sound*, McGraw-Hill, New York, 1948, pp. 172–213.

[4] Timoshenko, S., *Theory of Plates and Shells*, McGraw-Hill, New York, 1940, pp. 257–266.

[5] Wang, C. T., *Applied Elasticity*, McGraw-Hill, New York, 1953, pp. 276–282, 291–294.

[6] Synge, J. L. and Griffith, B. A., *Principles of Mechanics*, McGraw-Hill, New York, 1949, pp. 337–359, 418–447.

[7] Gray, A., Mathews, G. B., and MacRobert, J. M., *A Treatise on Bessel Function and their Applications to Physics*, MacMillan, London, 1931, pp. 9–27, 264–317.

[8] Carslaw, H. S., and Jaeger, J. C., *Operational Methods in Applied Mathematics*, Oxford University Press, London, 1953, pp. 348–352.

[9] Bowman, F., *Introduction to Bessel Functions*, Dover, New York, 1958, p. 135.

Satellite Networks for Global Coverage[1]

Frank W. Gobetz[2]

Abstract

This paper investigates the problem of establishing a network of satellites positioned such that at least one satellite is in view at all times from any point on Earth. In addition, the minimum number of satellites necessary to fulfill this condition is determined over a wide range of satellite altitudes. Because of their symmetry two orbit patterns are considered for this application: distributions of polar orbits, and arrangements composed of orbits placed perpendicular to the faces of the regular polyhedra. The former are found to be superior in all cases.

An analytical method is derived by which the optimum number of polar orbits for a given total number of satellites can be predicted. These optimized patterns are then considered in detail and operating altitudes are determined for systems consisting of from 6 to 100 satellites. An important consideration in this phase of the analysis is the proper synchronization of satellites in adjacent orbits as well as within each orbit. Previous solutions to this problem have disregarded such synchronization and are therefore unnecessarily redundant.

Six satellites appear to be the absolute minimum number of satellites for the two orbit patterns considered. The optimum distribution of six satellites yields complete coverage at all times at an altitude of only 5260 n. mi. Some consideration has been given to the problem of obtaining fewer satellites with other patterns, and a special pattern requiring only five satellites is described.

The calculated altitudes are found to compare favorably with ideal instantaneous coverage patterns composed of satellites placed at the vertices of regular polyhedra circumscribed about the Earth. Although these instantaneous arrangements comprise the best physical distributions of satellites for obtaining global coverage, the relative positions of the satellites cannot be maintained over a complete revolution and the patterns are thus impractical for an operating system.

Consideration is given to pattern stability and the polar array is shown to be an excellent choice from this standpoint. In addition, calculations are made to estimate the probable lifetime of each polar pattern, based on the satellite operating altitudes.

Symbols

r_0	Radius of Earth
s	Arc-length radius of satellite coverage
α	Central angle of satellite coverage
A	Spherical area of satellite coverage
h	Altitude
N	Number of orbits
n	Number of satellites per orbit
η	Total number of satellites
ψ	Angle formed by intersection of adjacent orbits
ψ_0	Spherical angle between orbits with co-directional motion
ψ'	Spherical angle between orbits with opposing motion
l	"Lead;" relationship between satellites in orbit with those in an adjacent orbit
x, y	Satellite positions measured from the North Pole
m	Satellite mass
λ	Lagrange Multiplier
Γ	Any integer
Z	Arc length defined in Fig. 13
L	Satellite lifetime
ϕ	Rate of rotation of the orbital plane about the Earth's rotational axis
i	Orbital inclination to the equatorial plane
T	Period of revolution
g_0	Gravitational acceleration at the Earth's surface
S	Satellite base area in drag calculation
h_δ	Satellite altitude based on restricted viewing angle

Introduction

Applications envisioned for artificial Earth satellites encompass such fields as meteorology, surveillance, navigation, and communication. A fundamental consideration inherent within each of these applications is the problem of distributing the satellites in such a way that every point on Earth is continuously in view of at least one satellite. The purpose of this investigation was to examine the problem of obtaining full-time, global coverage by a satellite network so conceived as to require a minimum number of satellites. Although no attempt is made to prove rigorously that the distributions discussed herein are optima, they nevertheless compare favorably with certain known ideal solutions.

Assumptions

All satellites are assumed to be at the same altitude and therefore only circular orbits are considered. The use of elliptic orbits apparently can not reduce the number of satellites required, and precession of the

[1] Submitted May, 1961.

[2] Research Engineer, United Aircraft Corporation Research Laboratories, East Hartford, Connecticut.

perigee of such orbits might make them undesirable for this application.

It was assumed that any number of satellites may be uniformly distributed within a prescribed orbit and that the orbits themselves may be attained precisely. Furthermore, all orbits in a given distribution contain the same number of satellites. This assumption is justified in part by Appendix I, in which arrangements with unequal numbers of satellites in each orbit are considered.

No stipulation is made as to which satellite will be visible to a given portion of the globe at a given instant. One or more satellites will be visible at all times although the particular satellites in view will change periodically. This simplification is an important one in that it allows the Earth's rotation to be neglected.

Analysis

The area on Earth visible to a satellite is a circle described by the locus of tangents from the satellite to the Earth. For example, in Fig. 1 the satellite "sees" the area between tangents SA and SB. Thus if s represents the arc radius of the small circle seen by the satellite,

$$\frac{s}{r_0} = \frac{\alpha}{2}, \tag{1}$$

$$\cos\frac{\alpha}{2} = \frac{r_0}{r_0 + h}, \tag{2}$$

therefore,

$$\frac{s}{r_0} = \cos^{-1}\left(\frac{r_0}{r_0 + h}\right). \tag{3}$$

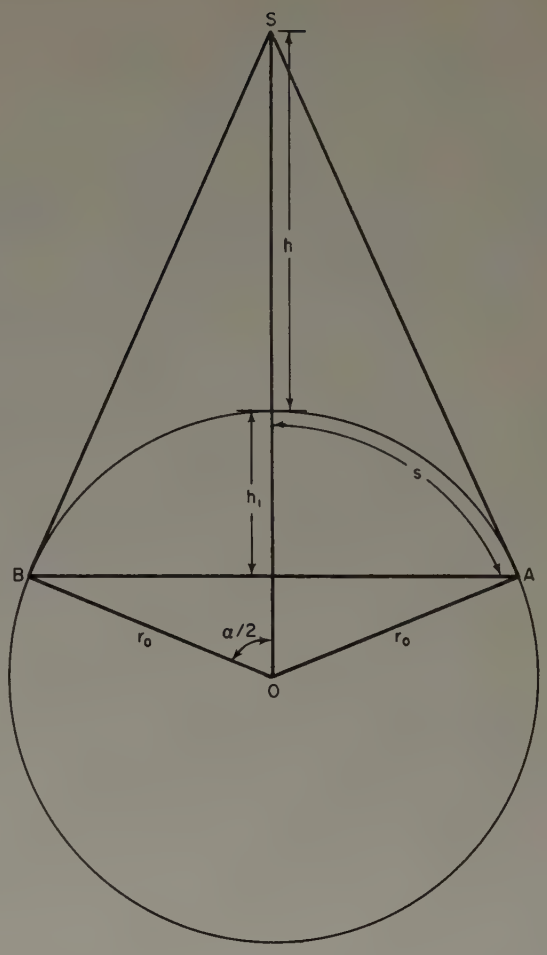

Fig. 1. Coverage of one Satellite

It is apparent that the nondimensional radius, s/r_0, of this small circle is a function of the satellite's altitude, and that the circle approaches a great circle as a limit when the altitude is infinite. Thus the two quantities s/r_0 and h/r_0 can be used interchangeably for a given Earth radius r_0. For convenience, s/r_0 will be used in most of the diagrams and in much of the explanatory discussion which follows. It is important to realize that the radius s/r_0 represents the altitude of a satellite in any diagram, even though the satellite itself may not be shown.

Since the assumption has been made that each orbit will contain the same number of satellites, it can be shown that the total number of satellites in any pattern may not be less than six (see Appendix II). That is, no orbital pattern can maintain global coverage for the entire period of revolution if less than six satellites are involved. A total of six could be achieved either by two satellites in each of three orbits or by three satellites in each of two orbits. Similarly a total of eight would entail two in four or four in two, etc.

With regard to the placing of the orbits, two arrangements are considered here: (1) an arbitrary number of orbits which intersect at two common points 180 deg apart; due to the Earth's oblateness perturbation the intersection points would normally have to be at the poles, but this will subsequently be discussed in more detail, and (2) orbits oriented so that they will be parallel to the faces of the regular polyhedra. For simplicity these two patterns will be referred to as "polar-type" and "polyhedron-type" distributions.

In determining altitude requirements for the various satellite patterns to be considered here it will be necessary to establish "critical" satellite orientations as design conditions. Obviously when the projections of two satellites on the Earth are in close proximity, the region between them is well covered. But if three satellite projections are regarded as the vertices of a spherical triangle ABC as in Fig. 3a, coverage of the midpoint 0 of the triangle requires that each satellite be at the altitude corresponding to s/r_0. The problem then, is one of finding the most serious "hole" opened up by the relative motions of the satellites, and establishing a minimum altitude requirement for its closure or coverage.

Polar Orbits

In analyzing the polar-type orbit arrangements the critical satellite orientations are easily ascertained from Fig. 2. Since the satellites are uniformly spaced in each orbit the most critical separation of satellites occurs where the orbits are most distant from each other. This

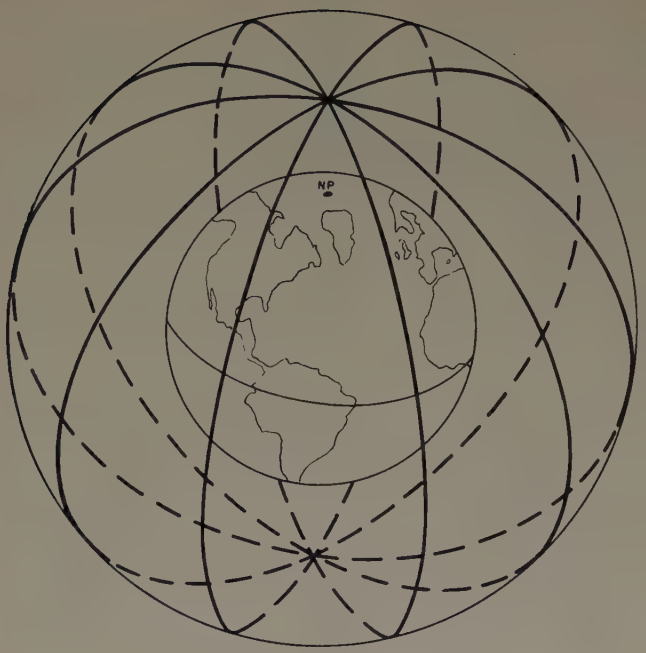

FIG. 2. Polar orbit paths

condition occurs 90 deg from the orbit intersection points, i.e., at the equator. Conversely, the most redundant coverage occurs over the intersection points where satellites move closer together.

Assuming that satellites in two adjacent orbits move in the same direction, the critical condition for covering the region between the two orbits is prescribed by some position of the spherical triangle ABC in Fig. 3a. Imagine a circle centered at 0 and passing through points A, B, and C. The radius, s/r_0, of this circle is maximized by moving the three vertices of the triangle in their respective orbits, always maintaining synchronization. This maximum value of s/r_0 represents the required altitude for coverage between two orbits in which the satellites move in the same direction. Unfortunately such synchronization is not possible between all adjacent orbits in a given polar-type pattern. The motions of satellites in at least two adjacent orbits will always have opposing sense and the altitude requirements thus imposed must be at least as bad as those found with co-directional motion.

The critical satellite orientation when the satellites in adjacent orbits move in opposite directions is one which exhibits symmetry about the centerline between orbits as in Fig. 3b. The determination of satellite altitude necessary to cover the region within the four points D, E, F, and G entails circumscribing a circle about the spherical quadrilateral formed by these points. The radius of the circumscribed circle is some s/r_0 which represents a higher altitude than that found when the motion was synchronized. Thus a condition such as that shown in Fig. 3b determines the operating altitude of a system of satellites in a polar-type orbit arrangement.

As an extension of this line of analysis, and for reasons which will subsequently be made clear, it is convenient to define an upper bound to the necessary satellite

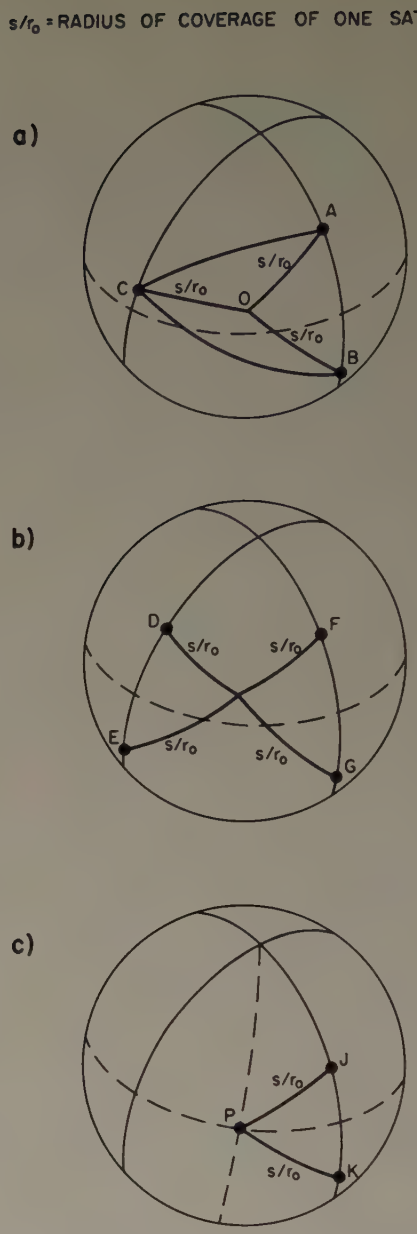

FIG. 3. Altitude requirements for polar orbits

altitude. This upper limit can be established by disregarding the positions of individual satellites and considering only a "band-width" covered by the satellites in any orbit. A simple determination of this bandwidth is illustrated in Fig. 3c. Two satellites J and K in the same orbit are in positions which are symmetric with respect to the "equator." If the magnitude of s/r_0 is such that both satellites can "see" point P, which lies on the equator and midway between the two orbits in the diagram, then the bandwidth extends one half the distance to the next adjacent orbit. Since each orbit contains the same number of equally-spaced satellites, the bandwidths are equal and the condition defined in Fig. 3c is more than sufficient to insure global coverage. That is, the magnitude of s/r_0 represents an altitude, $h_{\max}$, which is always equal to or greater than the previously defined operating altitude.

This upper altitude limit can be expressed as a simple function of the number of orbits, N, and the number of satellites per orbit, n. Similarly the altitude required to cover the region between two orbits whose satellites are synchronized is a simple function of N and n. Since the latter altitude is always less than the operating altitude and the former is always greater, a definite bound is placed on the desired operating altitude.

Determination of the operating altitude involves finding the midpoint of a spherical triangle whose sides are known. While this is not a particularly difficult operation, it involves the solution of several equations containing natural and inverse trigonometric functions which are too unwieldy to manipulate in an optimization operation. The use of the upper and lower bounds will thus be useful in the optimization procedure of Appendix III. From this analysis it becomes possible to predict, for a given total number of satellites, η, the number of orbits and satellites per orbit which will yield a minimum altitude solution. In this way the number of cases which must be examined in detail can be reduced, and at the same time limits can be established within which the desired altitude must fall.

As yet no stipulation has been made as to the relationship between satellites in adjacent orbits. Since the spacing of satellites within each orbit is uniform, satellites in adjacent orbits with co-directional motion will maintain the same phase relationship or "lead" to each other. A logical choice for good synchronization might be a lead of $\frac{1}{2}$ the spacing between satellites in a given orbit. Another distribution which was investigated is that in which the lead is $2\pi/n^2$. A procedure by which the best lead may be rapidly determined is outlined in Appendix IV.

Orbits Parallel to the Faces of Polyhedra

Configurations possessing a high degree of symmetry can be created if the satellite orbits are constructed parallel to the faces of the regular polyhedra. The orbits of each such network intersect at regular intervals thus describing intricate patterns of equilateral spherical polygons as shown in Fig. 4.

Placing a satellite at each intersection, the directions of motion can be adjusted so that collisions will never occur. Because the polygons formed by the orbits have equal sides, it can be concluded that the same sym-

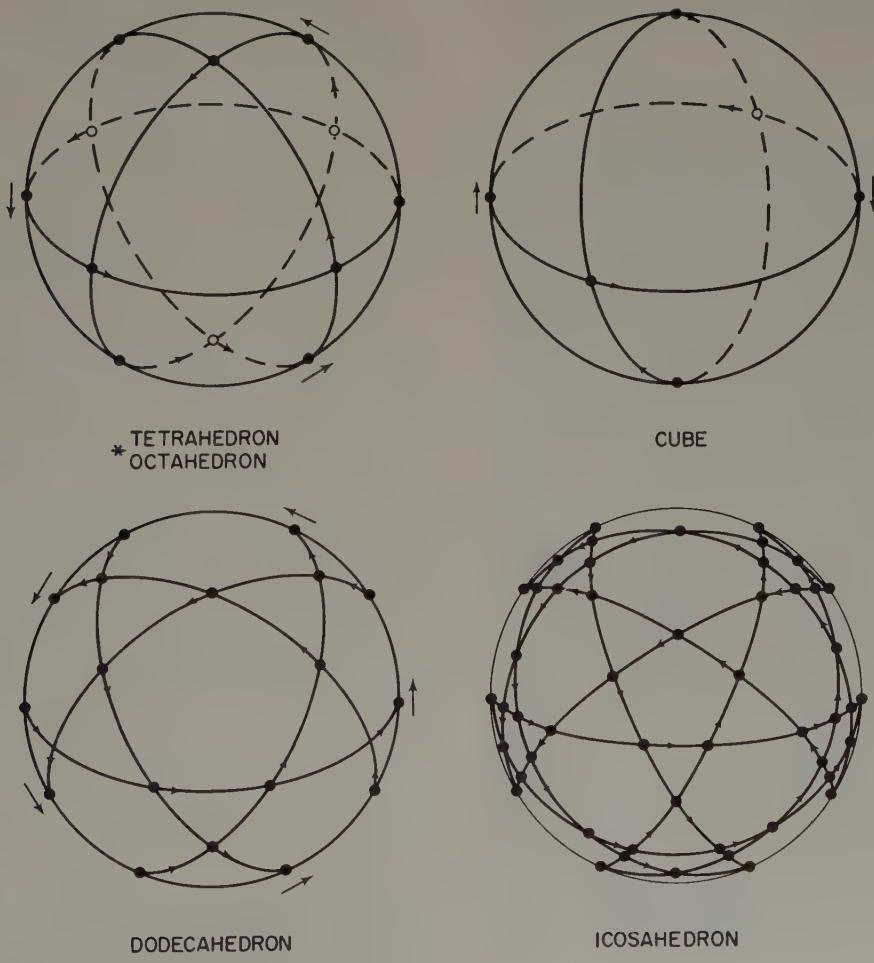

FIG. 4. Polyhedron orbits

metrical pattern will be repeated after every $2\pi/n$ radians of revolution. Furthermore the condition of maximum satellite separation must occur when each satellite is midway between two interesection points, and consequently it is this position which determines the proper altitude for global coverage.

Discussion of Satellite Distribution

Polar-Type Distributions

When the polar orbit distributions are optimized according to Appendix III, the two curves of Fig. 5 can be constructed. The optimum polar configurations are then easily determined for any number of satellites. For example, when $\eta = 6$ one could choose either $N = 2$ or $N = 3$, but since the former lies closer to the curves it is the better choice. Similarly, when $\eta = 32$, $N = 4$, and so on for all values of η.

Certain arrays are not valuable however, even though they represent the best solutions for given η's. One such case is $\eta = 16$, for which Fig. 5 predicts $N = 2$ rather than $N = 4$. But when the minimum satellite altitude

$$l = \frac{\pi}{n}$$

CIRCLES DENOTE USEFUL PATTERNS

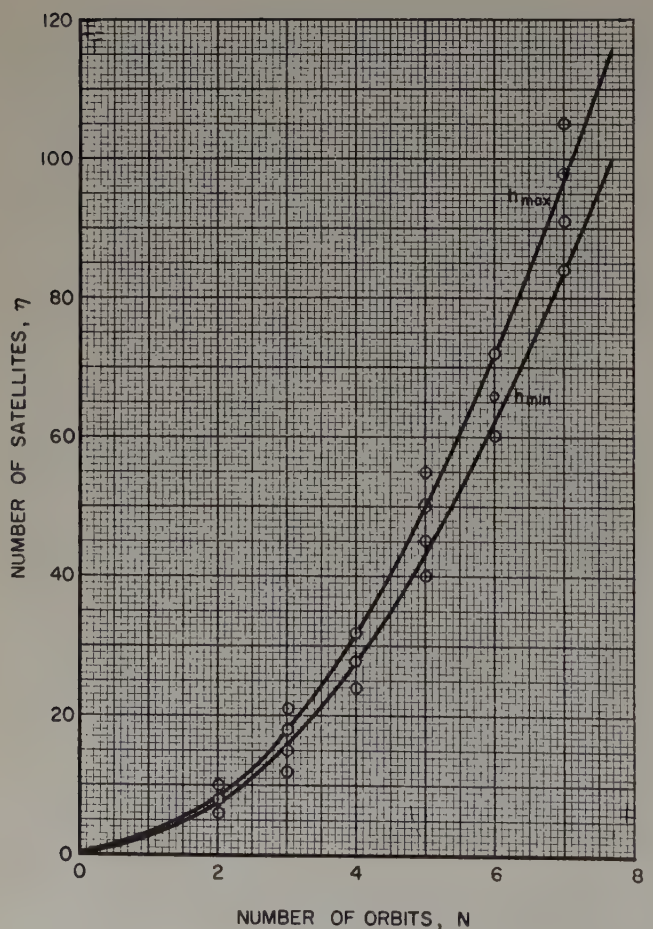

FIG. 5. Optimum polar configurations

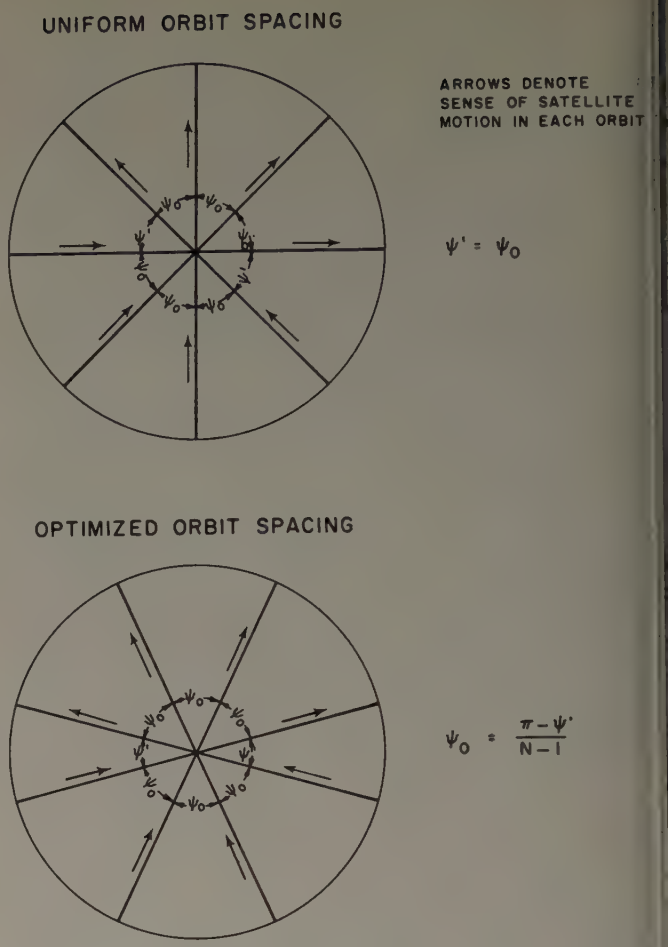

FIG. 6. Polar view of orbit spacing

is calculated, it is found to be greater than that for $\eta = 15$, $N = 3$. There are several such redundant cases, of which a few are $\eta = 16, 22, 25, 27, 30, \cdots$ etc.

The small circles in Fig. 5 represent optimized polar distributions which are not redundant. That is, each increase in the number of satellites entails a reduction, however slight, of the required altitude.

It has been previously stated that for polar arrays, the limiting satellite orientation is prescribed by opposing motion in adjacent orbits, and that only two such orbits need exist regardless of the total number of orbits involved. Obviously the altitude requirements can be reduced if the orbits are spaced so as to favor this particular pair, the technique being illustrated in Fig. 6.

When the orbit spacing is optimized in this manner, considerable reductions in altitude are achieved. Table I is a summary of the optimum polar distributions, including the appropriate leads l, optimized angles ψ'_{opt}. and the minimum altitudes, h_{opt}.

Polyhedron-Type Distributions

Table II summarizes the altitude requirements for the polyhedron arrangements. The column headed h_{min} refers to the symmetrical positions with satellites located at the intersections of the orbit paths, while the operating altitudes based on the worst satellite orientations are shown in the last column. None of the polyhedron

TABLE I

Characteristics of Optimized Polar Arrays

η	N	n	l (deg)	ψ'_{opt} (deg)	h_{opt} (n.mi)
6	2	3	40.0*	77.1	5260
8		4	45.0	83.3	2860
10		5	14.37*	86.7	2340
12	3	4	22.5*	46.3	1798
15		5	72.0	49.0	1194
18		6	10.0*	55.4	1037
21		7	25.7	53.6	840
24	4	6	30.0	37.2	732
28		7	7.34*	40.1	618
32		8	22.5	37.3	486
40	5	8	5.62*	31.4	425
45		9	20.0	28.9	326
55		11	16.35	30.5	275
60	6	10	18.0	21.7	241
72		12	15.0	24.3	201
84	7	12	2.5*	23.1	194
91		13	13.83	19.6	156
105		15	12.0	21.3	137
112	8	14	12.85	16.7	124

* Denotes $l = 2\pi/n^2$; all other leads are $l = \pi/n$.

TABLE II

Polyhedron Distributions for Full-Time Coverage

Figure	N	n	η	h_{min} (n.mi)	h (n.mi)
Cube	3	2	6	2515	∞
Tetrahedron	4	3	12	1425	3645
Dodecahedron	6	5	30	610	743
Icosahedron	10	9	90	565	609

TABLE III

Altitude Summary for Instantaneous Polyhedron Distributions

Figure	η	h/r_0	Description
Tetrahedron	4	2.0	Four equilateral triangles
Cube	8	0.732	Six squares
Octahedron	6	0.732	Eight equilateral triangles
Dodecahedron	20	0.237	Twelve regular pentagons
Icosahedron	12	0.320	Twenty equilateral triangles

The regular star polyhedra are excluded from this summary because only the convex polyhedra represent minimum altitude solutions. Star polyhedra have 12 vertices and all are inferior to the regular icosahedron from the standpoint of altitude. (see Refs. 6 and 7)

distributions was found to be better than a polar arrangement requiring the same total number of satellites.

Comparison of Distributions

In order to properly evaluate the merit of the solutions arrived at here, a comparison was made with the absolute minimum altitudes as defined in Appendix V. This appendix deals with a fictitious case in which the satellite coverage areas are "distorted" from their circular shape in order to reduce the overlap to zero.

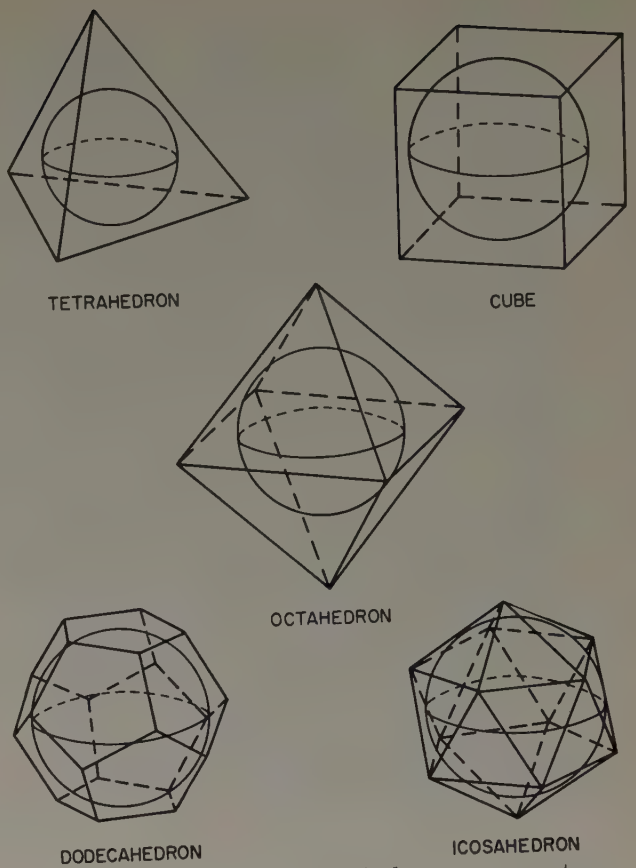

FIG. 7. Instantaneous polyhedron arrangements

While these solutions are not physically attainable their usefulness lies in the fact that they represent a condition of zero redundancy.

Another arrangement useful for purposes of comparison is one composed of satellites placed at the vertices of a regular polyhedron circumscribed about a spherical Earth. Due to its symmetry this configuration is the best possible physical solution but provides only periodic coverage. That is, the orbits cannot be so placed as to retain the positions of the satellites relative to each other and thereby maintain continuous coverage. A summary of the altitude requirements for each of the five regular polyhedra is provided in Table III and Fig. 7 depicts the satellite arrays.

Figure 8 compares these two idealized cases with the real solutions of Tables I and II and illustrates several interesting points. Both the smooth curve and the lower step curve approach an infinite number of satellites when the altitude is close to zero. However as the altitude is increased the ideal curve smoothly approaches two satellites while the real solution remains at six until infinity is reached, at which time the final step in the curve goes to two satellites. As expected, the solutions corresponding to the instantaneous regular polyhedra fall in between the two curves. Due to the conclusions of Appendix I there is no advantage to be gained by filling the orbits with unequal numbers of satellites if $\eta > 5$, and therefore the lower step curve of Fig. 8 represents the best results obtained in this

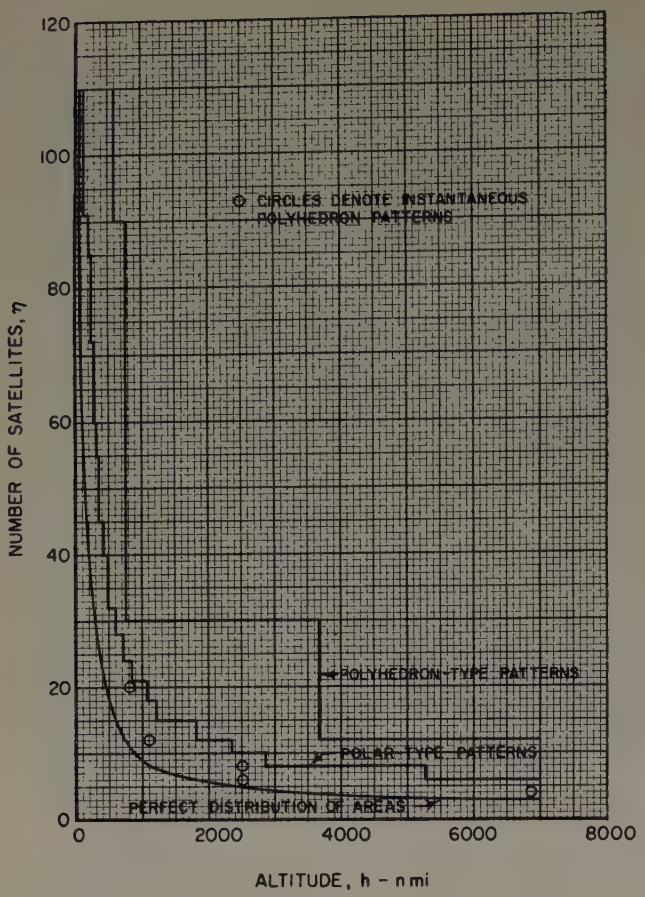

FIG. 8. Comparison of real solutions with instantaneous polyhedron arrays and perfect distribution of areas.

analysis. As discussed in Appendix I, the case of five satellites in three orbits probably does not constitute a better solution, since the operating altitude is several times that required for six satellites.

The satellite operating altitudes discussed thus far have been based on the assumption that a satellite can "see" an area on Earth described by the locus of tangents from the satellite to the Earth as in Fig. 1. Even if the Earth were a smooth sphere this assumption would be an optimistic one since the Earth's atmosphere causes a refraction of light and radio waves. This effect is most pronounced near the "twilight" region of the satellite coverage area because an observer in this region must look through a greater thickness of atmosphere in viewing the satellite.

To compensate for this effect the right angles SA0 and SB0 in Fig. 1 must be increased by some angle, δ. As a result, the satellite altitude must be increased to a value greater than h in Fig. 1. The resulting altitude, h_δ, is plotted against the ideal operating altitude h in Fig. 9. Therefore, if the angle δ is assumed to be zero, h and h_δ are equivalent, but for $\delta > 0$, $h_\delta > h$.

Network Stability

The solutions labeled "optimum" share the common feature that they depend upon a complex synchronization of satellites and orbits. Such precision is to no avail however, if this synchronization cannot be maintained for reasons of stability. At least a qualitative consideration of perturbations to which the satellites will be subjected must therefore be undertaken.

Satellite perturbations arise from three main sources: the Earth's oblateness, the atmosphere, and third bodies such as the Sun and Moon. Of these, the first two are by far the most significant (Refs. 3 and 4), and accordingly third-body effects will be disregarded here. Solar radiation pressure could also cause a substantial perturbation if the satellite surface area to mass ratio were unusually high, as in the case of the Echo balloons. Since this study is not directly concerned with passive communication satellites however, this perturbation will be neglected.

The oblateness of the Earth has a threefold effect on satellite orbits.

1. Rotation of the major axis within the orbital plane
2. Reduction of the period
3. Rotation of the orbital plane about the Earth's axis

Since only circular orbits have been considered here rotation of the major axis is not pertinent. Effects on period of revolution cannot be disregarded, but these may be compensated by adjustment of the altitude of the satellite orbit.

Rotation of the orbital plane is predicted by the expression (see Ref. 2)

$$\phi = 10 \left(\frac{r_0}{r_0 + h} \right)^{3.5} \cos i \, \frac{\text{deg}}{\text{day}} \qquad (4)$$

It can be seen that the rate of rotation is directly related to orbital inclination, i. Thus, polar orbits are unaffected in this respect, and circular orbits of equal inclination are all affected identically. This is quite important in the present application, since the positions of satellites relative to each other, are not relative to the Earth, are of primary concern.

With this in mind, a careful re-examination of the two-orbit case ($N = 2$) reveals an important exception. Regardless of the location of the orbit intersection points, the two orbits can have equal inclinations to the equatorial plane. Therefore if the intersection points are not at the poles, the orbits rotate at the same rate about the Earth's axis and satellite coverage remains unchanged. This is a characteristic of only the two-orbit solutions, however, and is not true for more than two orbits.

Satellite perturbations due to the Earth's atmosphere result in a decrease of satellite altitude which advances rapidly as the dense region near the Earth is approached. Thus a pattern consisting of 78 satellites in a very low orbit has a lifetime of only a few weeks, while a six-satellite pattern at several thousand miles altitude can be expected to remain in orbit indefinitely. Satellite lifetimes are plotted against η in Fig. 10. In the preparation of this graph the method of Ref. 2 was followed and numerical data was obtained from Refs. 4 and 5.

It can be concluded with reasonable assurance from

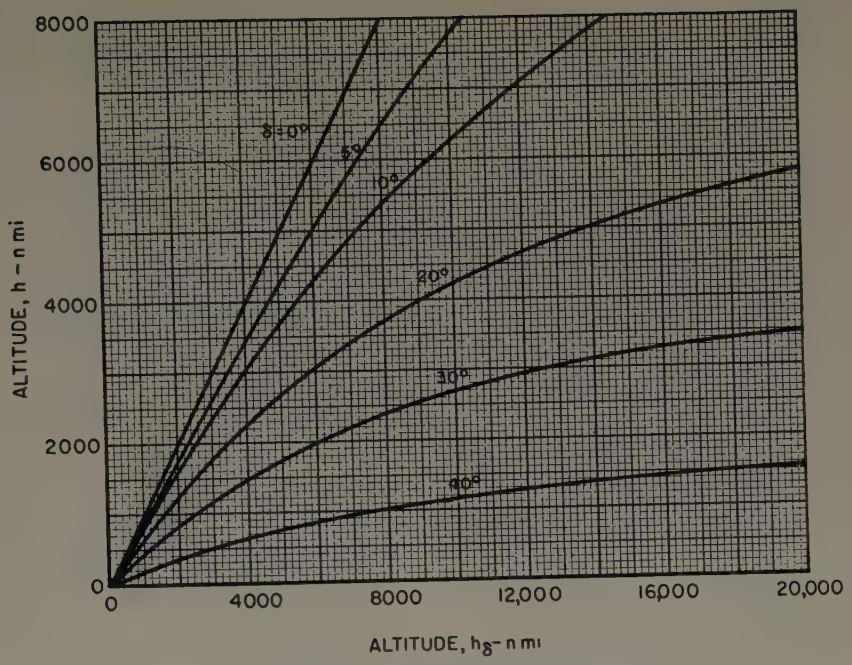

FIG. 9. Effect of restricted viewing angle

S/m = 0.01 ft^2/lb

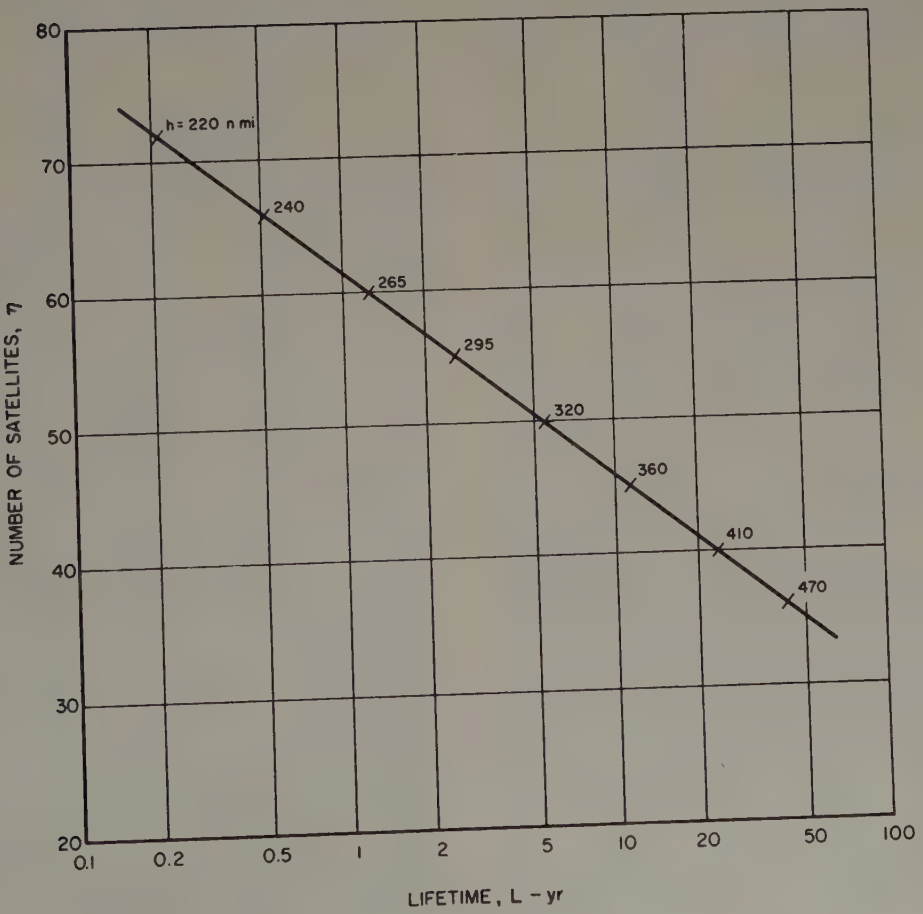

FIG. 10. Pattern lifetimes

TABLE IV

Unequal Numbers of Satellites in Each Orbit

N	a	b	c	d	η
2	2	4	—	—	6
	3	4	—	—	7
	3	5	—	—	8
	3	6	—	—	9
	4	5	—	—	9
	4	6	—	—	10
	5	6	—	—	11
	5	7	—	—	12
3	1	2	2	—	5
	2	2	3	—	7
	3	3	4	—	10
	2	3	3	—	8
	3	4	4	—	11
4	2	2	3	3	10
	3	3	4	4	14

the preceding stability considerations that polar orbit patterns are the best practical arrays for global coverage. It had been previously stated that symmetry alone was a good argument for the use of polar distributions, but the inherent stability of these orbits constitutes another important feature not likely to be found in other arrays.

Concluding Remarks

An important consideration in this analysis was the proper synchronization of satellites in adjacent orbits as well as within each orbit. Previous attempts at solution of the global coverage problem (Refs. 1 and 8), have disregarded such synchronization and are therefore unnecessarily redundant.

Since only two orbital patterns have been investigated, the results presented here cannot be considered as absolute optima. However, a fundamental premise of this analysis has been that the desired optima will possess a high degree of symmetry. Figure 8 shows clearly that the instantaneous polyhedron distributions, which are extremely symmetrical, fall close to the step curve, and one can therefore conclude that if improvements are to be made they must at best be small.

Bibliography

1. Vargo, L. G., "Orbital Patterns for Satellite Systems," American Astronautical Society Preprint No. 60–48, January 1960.
2. King-Hele, D. G., and R. H. Merson, "Satellite Orbits in Theory and Practice, Journal of the British Interplanetary Society," Vol. **16**, No. 8, pp. 446–471, July–August 1958.
3. Moulton, F. R.: *Celestial Mechanics*, pp. 321–362, MacMillan Company, New York, 1958.
4. Sterne, Theodore E., "High Altitude Atmospheric Density," The Physics of Fluids, Vol. **1**, No. 3, pp. 165–170, May–June 1958.
5. Krasovskii, V. I., "Results of Scientific Investigations Made by Soviet Sputniks and Cosmic Rockets," *ARS Journal*, Vol. **30**, No. 1, pp. 27–33, January 1960.
6. Coxeter, H. S. M., *Regular Polytopes*, Methuen & Co. Ltd., London, 1948.
7. Cundy, H. Martyn and A. P. Rollett, *Mathematical Models*, Clarendon Press, Oxford, 1951.
8. Luders, R. David, "Satellite Networks for Continuous Zonal Coverage," *ARS Journal*, Vol. **31**, No. 2, pp. 179–184.

Appendix I

Unequal Numbers of Satellites in each Orbit

When each orbit is required to contain the same number of satellites, a restriction is placed on the total number of satellites in the pattern. Thus if

$$\eta = nN \tag{5}$$

where

$$n \geqq 2$$

$$N \geqq 2,$$

distributions containing $\eta = 5, 7, 11, 13, 17, 19$, etc. are excluded. In order to obtain such values of η it is necessary to relax this restriction.

The cases investigated in detail are summarized in Table IV in which $a, b, c, d, \cdots$ represent the number of satellites in each orbit where $a \leqq b \leqq c \leqq d \cdots$ etc. Each distribution was examined only until it could be shown to be inferior to a corresponding case in Table I or Table II and therefore the altitude requirements are not listed.

It should be pointed out that these results are inconclusive in one respect. When adjacent orbits contain unequal numbers of satellites and the satellites in each orbit are spaced uniformly, there is a considerable latitude of choice in selecting an appropriate lead between satellites in adjacent orbits. Due to the trial-and-error nature of this phase of the analysis, only a few possibilities could be examined in each case, and the results are therefore uncertain. In most cases however, the altitudes calculated were prohibitively great and could not be made competitive by any juggling of the satellite leads.

Only one instance was found in which it was advantageous to fill the orbits with unequal numbers of satellites. This is the five-satellite distribution which is composed of three orbits. The method of coverage in this case takes advantage of the basic asymmetry in any pattern composed of an odd number of satellites. In Fig. 11a two satellites are shown in the same orbit in such relative positions that they leave a lune-shaped area uncovered. This uncovered area, only half of which is visible in the diagram, rotates in a clockwise manner as the satellites revolve in the plane of the paper. Because two finite-altitude satellites cannot cover a great circle (see Appendix II), the center of the hemisphere in Fig. 11a is within the shaded area. Two satellites can also be placed in an orbit which is perpendicular to the first as shown in Fig. 11b. Again only half of the region left uncovered by these two satellites is visible in the diagram, and the sense of rotation of the shaded area proceeds from the lower part of the diagram toward the top.

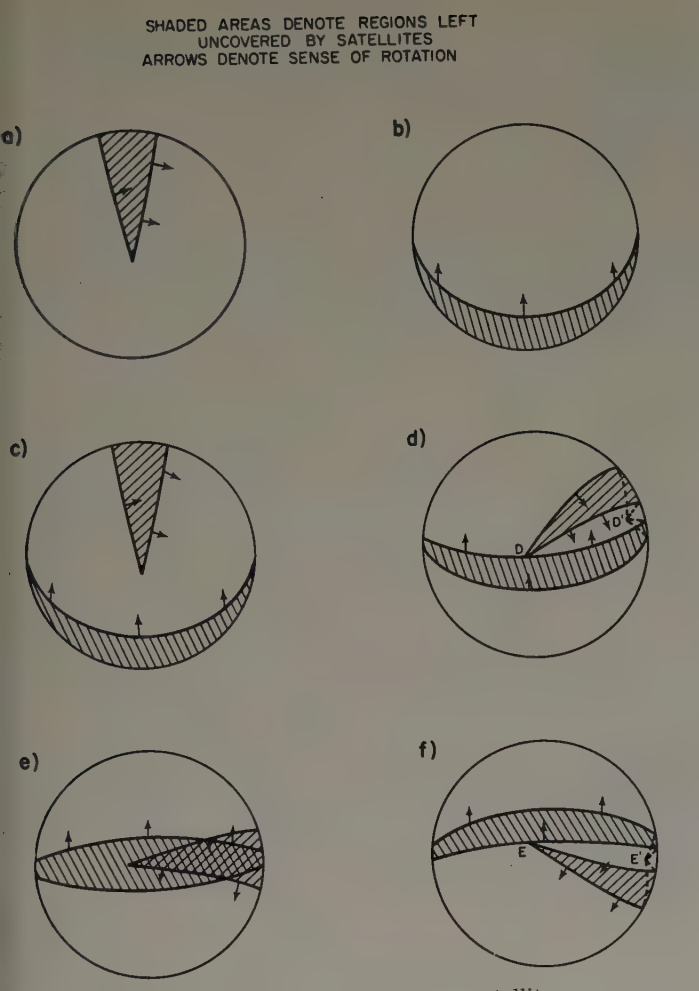

FIG. 11. Global coverage by five satellites

If these two cases are now combined as in Fig. 11c, the entire sphere is covered for the instantaneous satellite positions shown in the sketch. When the satellites are allowed to revolve in their assigned directions, however, the shaded regions soon meet, indicating a loss in coverage. Thus in Fig. 11d coverage is first interrupted at point D and at another point D′ which is in the hemisphere hidden in Fig. 11d. Note that the arc length D-D′ is only slightly greater than 90 deg. The uncovered (cross-hatched) region then spreads until in Fig. 11e its area reaches a maximum. Then, as in Fig. 11f it decreases in size to two points E and E′ which are analogous to D and D′ in Fig. 11d. By visualizing the further rotation of the satellites it is not difficult to see that the uncovered regions arise at symmetrical positions on the globe (180 deg apart). Therefore a single satellite could be placed in a third orbit and synchronized in such a way that it could cover these regions as they arise.

Due to the complexity of this system a general operating altitude has not been determined, although the altitude must exceed that for an instantaneous tetrahedron distribution, which is two Earth radii. One special case was examined in detail and the result is probably close to the minimum altitude which can be obtained with a five satellite network. In this system the plane of the fifth satellite's orbit would be viewed on edge in Fig. 11 and the satellite would pass from left to right on the visible side. For this particular case the minimum operating altitude was found to be 13 Earth radii or 44,720 n. mi.

Appendix II

Minimum Number of Satellites

In the body of this report the assumption is made that each orbit will contain the same number of satellites, n. As a consequence of this assumption distributions containing less than a total of six satellites cannot achieve full-time global coverage. The following is a proof of this hypothesis.

1. Consider two satellites at infinite altitude as shown in Fig. 12a. Each satellite is in view of a great circle (i.e., a hemisphere) and, together the two satellites maintain continuous global coverage.

2. If these two satellites are now brought in slightly from infinity to some very great altitude a condition such as shown in Fig. 12b results. A band of width ϵ remains uncovered because for $h < \infty$ a satellite cannot cover a great circle. In fact this statement can be extended to include any two satellites. That is, two satellites at finite altitudes cannot cover a great circle.

3. Consider a large number of finite-altitude satellites in a single orbit as in Fig. 12c. Two circles of diameter ϵ remain uncovered. It can therefore be concluded that for $h < \infty$, at least two orbits are required for global coverage.

4. Because the two uncovered regions in Fig. 12c are 180 degrees apart, they lie on a great circle. In step 2 it was shown that at least three satellites are required to cover a great circle. Therefore, if $N = 2$, $n \geqq 3$, and $\eta \geqq 6$. Since $N > 1$ by step 3, it can be concluded that $\eta \geqq 6$.

Appendix III

Polar Pattern Optimization

When the total number of satellites in a polar orbit pattern is specified, the exact distribution of satellites and orbits is not uniquely determined. For example, an arrangement consisting of twelve satellites may be composed of two, three, four, or six orbits with six, four, three, or two satellites in each. When the number of satellites, η, becomes large the numerous possible distributions become quite prohibitive to any trial-and-error approach. Clearly, a technique is desired by which this choice can be made less laborious.

Due to the complex synchronization of satellites in adjacent orbits, the desired altitude requirements are not reducible to a simple analytical form. However, the altitude limits h_{max} and h_{min} may be expressed as relatively simple functions of N and n. If these limits are minimized holding η constant, the optimum distributions should be defined by the solutions thus obtained.

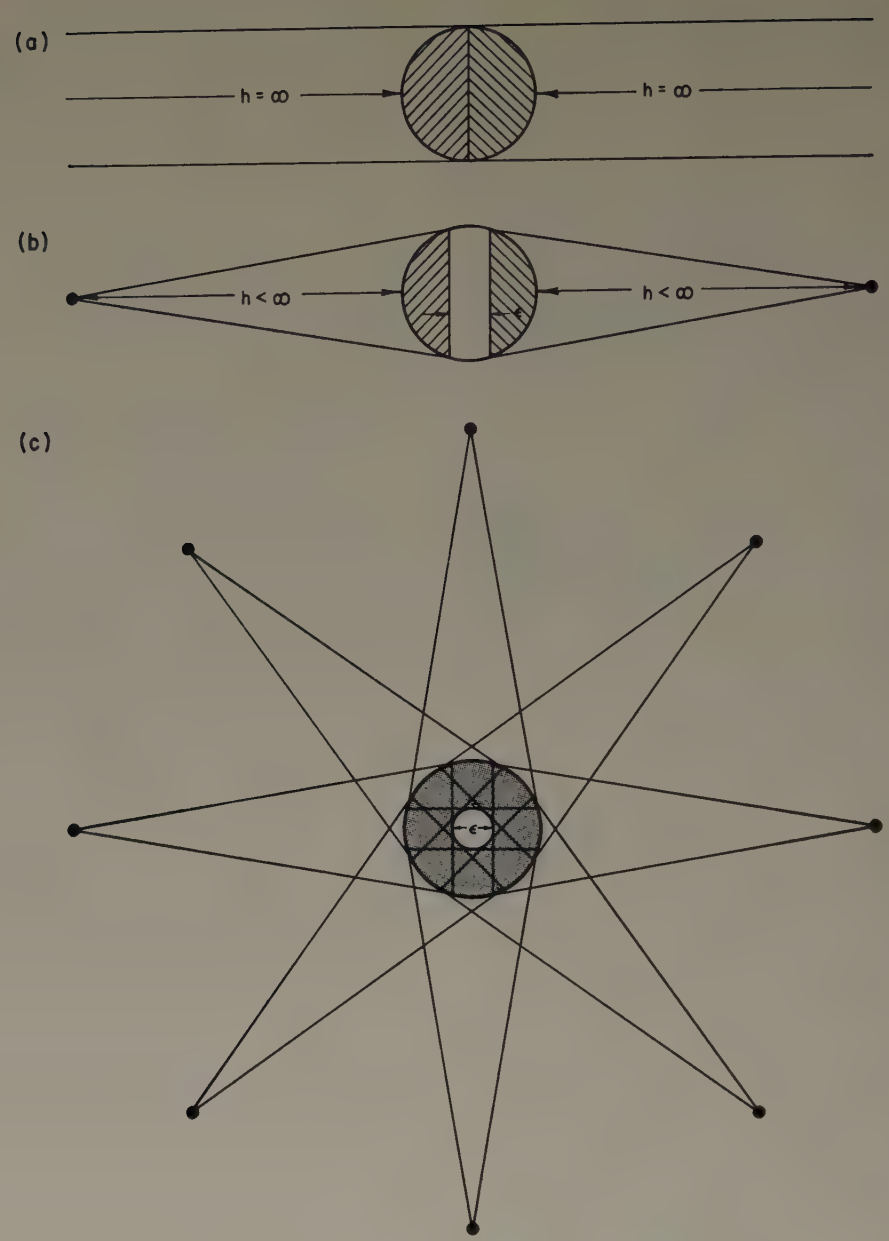

FIG. 12. Proof of minimum number of satellites

Referring to Fig. 3c, the maximum altitude limit can be expressed by:

$$h_{\max} = \frac{1}{\cos \frac{s}{r_0}} - 1 = \frac{1}{\cos \frac{\pi}{n} \cos \frac{\pi}{2N}} - 1. \quad (6)$$

If the total number of satellites is held fixed,

$$\eta(n, N) = \text{Constant} \quad (7)$$

$h_{\max}$ can be minimized as follows:

$$\frac{\partial h_{\max}}{\partial n} + \lambda \frac{\partial \eta}{\partial n} = 0 \quad (8)$$

and

$$\frac{\partial h_{\max}}{\partial N} + \lambda \frac{\partial \eta}{\partial N} = 0. \quad (9)$$

Solutions of Eqs. (6) through (9) results in

$$\frac{\tan \frac{\pi}{n}}{n} = \frac{\tan \frac{\pi}{2N}}{2N} \quad (10)$$

therefore,

$$n = 2N. \quad (11)$$

Lüders has also obtained this result in Ref. 8.

A similar procedure leads to optimization of $h_{\min}$. It is necessary in this case to make the assumption of a $\frac{1}{2}$ lead between satellites in adjacent orbits, but this restriction is not severe. Thus satellite C in Fig. 3a is assumed to lie on the equator and

$$h_{\min} = \frac{1}{\cos \frac{s}{r_0}} - 1 = \frac{1}{\cos \frac{\pi}{n} \cos \left(\frac{\pi}{N} - \frac{s}{r_0} \right)} - 1. \quad (12)$$

It is expedient here to rearrange Eq. (12) to a different form. Thus

$$F(n, N) = \tan \frac{s}{r_0} = \frac{\sec \frac{\pi}{n} - \cos \frac{\pi}{N}}{\sin \frac{\pi}{N}}. \qquad (13)$$

The parameter $F(n, N)$ may be minimized in place of $h_{\min}$ since it is the arc radius s/r_0 which is really being optimized in either case. As before

$$\frac{\partial F}{\partial n} + \lambda \frac{\partial \eta}{\partial n} = 0, \qquad (14)$$

and

$$\frac{\partial F}{\partial N} + \lambda \frac{\partial \eta}{\partial N} = 0. \qquad (15)$$

Solution of Eqs. (13) through (15) with the constraint of Eq. (7) yields

$$\frac{N}{n} \sin \frac{\pi}{N} \tan \frac{\pi}{n} = \cos \frac{\pi}{n} - \cos \frac{\pi}{N}. \qquad (16)$$

Although this expression cannot be solved explicitly for either variable, a graphical solution is easily obtained. This graphical solution together with the solution for $h_{\max}$ in Eq. (11) is shown in Fig. 5 as curves of η vs N. In order to determine N for some particular η, it is necessary only to choose the N closest to the curves.

Appendix IV

Optimum Lead

The lead is of considerable importance because it affects not only $h_{\min}$, but also the critical satellite orientation illustrated in Fig. 3b. The least desirable condition is one which possesses symmetry about the equator as well as about the centerline between orbits. A judicious choice of the lead can prevent this limiting orientation, thus directly lowering the required altitude for complete coverage.

Consider the pattern pictured in Fig. 13a where one complete orbit and half of each remaining orbit are shown. For convenience the satellite placed at the upper intersection belongs to the completed orbit.

The desired limiting condition involves the shaded area since the orbits bounding this region have opposing motion. If the lead is l, a satellite in the interior orbit must be at

$$y = l(N - 1), \qquad (17)$$

where y is measured from the upper intersection point. Similarly a satellite in the exterior orbit is at

$$x = \frac{2\pi}{n}. \qquad (18)$$

If these satellites are now allowed to move until they are in the adjacent positions depicted in Fig. 13b,

$$x = y = \frac{l(N - 1)}{2} + \frac{\pi}{n}. \qquad (19)$$

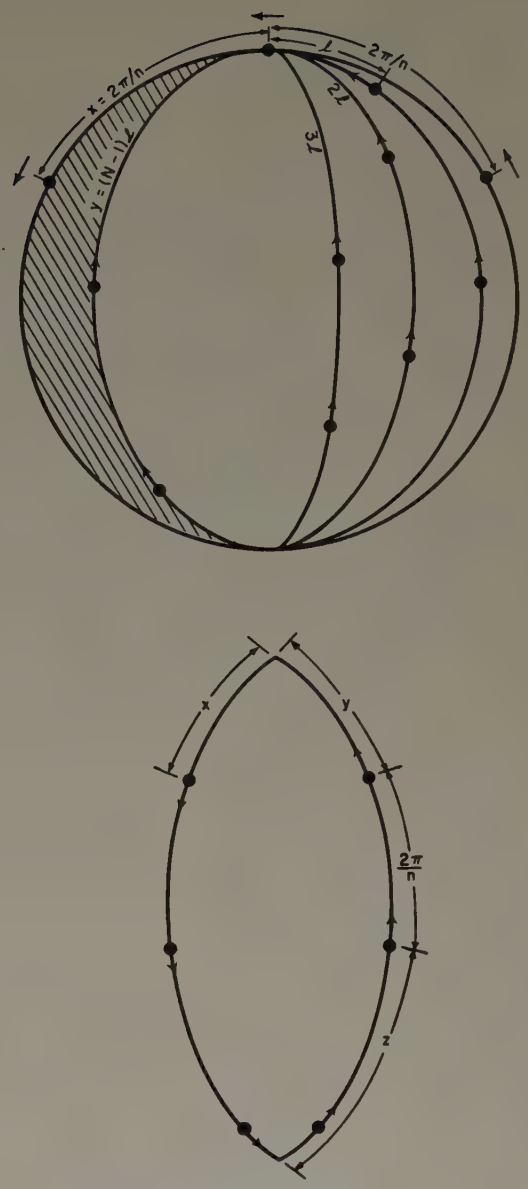

FIG. 13. Determining the optimum lead

Also

$$Z + y + \frac{2\pi}{n} = \pi. \qquad (20)$$

But such a condition will occur after each π/n radians of revolution. Thus

$$y = \frac{l(N - 1)}{2} + \frac{\pi}{n} \pm \Gamma \frac{\pi}{n} \qquad (21)$$

where

$$|\Gamma| = 0, 1, 2, 3, \cdots \qquad (22)$$

Given n, N, and l it remains now to determine the value of Γ for which the satellites assume positions most nearly symmetric with respect to the equator. This condition is represented by a minimal value of the quantity

$$|Z - y| = \frac{\pi}{n}\left(n - 4 - 2\Gamma - \frac{ln}{\pi}(N - 1)\right) \qquad (23)$$

This, of course, is subject to the condition that Γ take on only integer values.

The worst possible lead in any case would be one for which $|Z - y|_{\min} = 0$ and conversely the optimum lead is one for which $|Z - y|_{\min}$ is a maximum.

Equation (23) is perfectly general in that it facilitates the examination of any lead and the determination of a characteristic value of the quantity $|Z - y|_{\min}$. To compare two leads, $|Z - y|_{\min}$ is calculated for each. The larger value corresponds to the better choice of leads, in the sense that it will yield a lower satellite operating altitude.

A wide choice of leads was not considered in arriving at the results shown in Table I. It was found that leads of π/n and $2\pi/n^2$ were close to optimum in several cases and in all subsequent calculations only these two leads were considered. The relatively small reductions in satellite altitude which may be obtained by investigation of other leads does not appear to justify the extensive calculations which would be required.

Appendix V
Perfect Distribution of Area

As shown in Fig. 1 the area visible to a satellite is circular in shape. Therefore if the Earth is to be covered entirely, considerable overlap of adjacent circles must take place, resulting in redundant coverage. The area of the zone cut out on the sphere by a single satellite is determined as follows: (Referring to Fig. 1)

$$\cos\frac{\alpha}{2} = \frac{r_0 - h_1}{r_0} = \frac{r_0}{r_0 + h}, \tag{24}$$

$$h_1 = \frac{r_0 h}{r_0 + h}, \tag{25}$$

therefore,
$$A = 2\pi r_0 h_1 = \frac{2\pi r_0^2 h}{r_0 + h}. \tag{26}$$

If the areas covered by each satellite were perfectly distributed with no overlap, the number of satellites required for global coverage would be

$$\eta = \frac{4\pi r_0^2}{A} = 2\left(1 + \frac{r_0}{h}\right). \tag{27}$$

Book Reviews

Review of Methods of Celestial Mechanics, by Dirk Brouwer and G. M. Clemence.

At the present time the interest in celestial mechanics, particularly in its application to space vehicles, is in the midst of a remarkable revival. In spite of this rebirth of interest, the new textbooks available have been few in number; and those that were available were weak—weak particularly with respect to emphasis on numerical applications. It is, therefore, with considerable delight that students of celestial mechanics welcome the new book by Dirk Brouwer and Gerald M. Clemence, entitled *Methods of Celestial Mechanics* (Academic Press, 1961, $10.50).

Although *Methods of Celestial Mechanics* does not really consider applications to space vehicles, i.e., astrodynamics (only Section 14 of Chapter XVII specifically treats Earth satellites), and concerns itself very little with orbit determination, it does lead the reader clearly and logically down the road to a thoroughgoing understanding of orbit prediction techniques. Both special and general perturbations are treated in considerable detail, as are the subjects of aberration and the method of least squares.

In Chapter I are derived the equations of relative motion and the *vis-viva* equation, as well as a number of other relationships useful in the study of elliptical motion. Section 21 introduces matrix notation and calls attention to the original work of Cayley, while the exceedingly useful f and g series are developed in Section 30.

Fourier series and harmonic analysis are presented in Chapter II, and a useful definition of general perturbations is included; i.e., ". . .expansion in Fourier series where the arguments are retained as literal quantities while the coefficients are expressed as numbers." In order to lay down a fundamental background for the analysis of orbits about an aspherical planet, Chapter III develops the gravitational attraction between bodies of finite dimensions. In particular pages 125 ff. exhibit formulas that are well suited for the analysis of the motion of an object about a triaxial ellipsoid (such as the Moon).

Chapter IV discusses the algebraic, tabular, and graphical methods for the representation of data and then proceeds with an excellent presentation of polynomial interpolation formulas. Lagrange's (including a generalization), Everett's, Bessel's, and Newton's (forward and backward difference) formulas are all discussed and well illustrated with numerical examples. The way is thereby logically prepared for numerical integration. Here they show, among other things, that the "probable error of numerical integration after n steps is $0.1124n^{3/2}$ (in units of the last decimal), ..." Certain useful symbolic operators are also defined in this chapter.

In Chapter V Cowell's method is applied to the close approach of Ceres on January 6, 1941 and a derivation and numerical example of Encke's method are given as well. With respect to the choice of special vs. general perturbations the following words bear considerable significance: "It cannot be known yet whether either method will gain at the expense of the other (numerical vs. analytical), but it is certain that the practical celestial mechanician will always profit by the use of a judicious combination of numerical and analytic methods."

The following chapter includes a lucid discussion and definition of planetary and stellar aberration. The next chapter (Chapter VII) then discusses the comparisons of observation with theory, dwells briefly on a combination of special and general perturbations, and concludes with good discussions of precession, nutation, and geocentric parallax. Of some note here is the equation for sidereal time referred to the equinox of 1950.0 found on page 208.

A thoroughgoing discussion of error theory, including accidental and systematic errors, Gaussian distribution, singularities, and ways and means to solve by least squares are

liscussed in Chapter VIII. Chapter IX comes as close as ny to the subject of orbit determination in its treatment of differential correction, although the reader is really given no nsight into the question of the generation of a preliminary orbit. This problem is presently one of the central ones in astrodynamics, if not modern celestial mechanics.

General integrals and equilibrium solutions are considered n Chapter X, where the force function, F, is introduced. The Jacobian integral, Tisserand's criterion, surfaces of zero relative velocity, and the Lagrangian points are all discussed. In this latter regard they utilize the standard method for the treatment of small oscillations in mechanical problems; i.e., they develop linear differential equations with constant coefficients by expanding the coordinates of an object in the neighborhood of a libration point in a Taylor's series and retain only the first-order terms. For a Trojan minor planet they find that the motion about the equilibrium position consists of a libration with a period of about 147 years and a more rapid superimposed oscillation with a period slightly in excess of the period of Jupiter (about 11 years).

The chapter on variation of arbitrary constants is a rather involved one and requires careful study. Lagrange's brackets, Delaunay's method, etc., are all discussed. The lunar theory, which is usually understood to be "the problem of finding the Moon's motion under the gravitational attraction of the Earth and Sun, all three bodies being treated as mass points," is thoroughly investigated in Chapter XII. It was of particular interest to the present author that Brouwer and Clemence note that Brown's tables of the Moon don't reflect the best values of the lunar parallax and mass. The evection (largest periodic perturbation in the Moon's longitude—about 1° 16′), the annual equation, and the parallactic inequality are all discussed. It is also of interest to repeat the comment that none of the other classical methods can compete with Rabe's dynamical method for evaluating the solar parallax. Hill's lunar method and Brown's differential correction procedure then conclude the chapter.

The perturbations of the coordinates, which seems rather analogous to Encke's method and effectively and interestingly utilizes the *vis-viva* equation, is the subject of Chapter XIII. This same chapter introduces the reader to Hansen's device and Brouwer's special perturbations method. Hansen's method (Chapter XIV) is felt to be the best procedure for treating first-order perturbations because its rapidly converging series is applicable to higher orders of i and e than most other methods. Furthermore, "Hansen appreciated the advantages of applying all perturbations of both long and short period to the . . . mean anomaly." An interesting treatment of mean and osculating elements can be found in Section 8 of this chapter.

Chapter XVI demonstrates how to obtain secular terms by utilizing Lagrange's method. Section 6 specifically discusses Jacobi's determinant solution in which one reduces the off-diagonal terms to zero (a procedure that has been much used in differential correction of late). Other topics such as the limits on e and $\tan i$ in the solar system and families of minor planets are introduced here also. The concluding chapter on canonical variables investigates infinitesimal contact transformations and other aspects of Hamiltonian mechanics. Here we find the only application to artificial satellites. Of particular interest in this regard is an exposition of Hori's procedure, in which he removes the singularity for i near 63° 26′ by expanding in terms of the square root of the second harmonic.

From cover to cover *Methods of Celestial Mechanics* is an excellent book. It is not truly a book that one without some experience or prior introduction to celestial mechanics or astrodynamics can easily fathom—but it does evolve each topic from first principles and should be comprehendible to an advanced graduate student from other scientific disciplines. Certainly for the graduate celestial mechanic or astrodynamicist it is a book well worth studying and represents a notable contribution to the field.

Robert M. L. Baker, Jr.
Department of Engineering,
University of California,
Los Angeles, and
Lockheed California Company
Astrodynamic Research Center

Review of The Physical Principles of Astronautics, by Arthur I. Berman.

The Physical Principles of Astronautics by Arthur I. Berman (John Wiley and Sons, N. Y., $10.50) is a book designed to give the college student a concise and thorough exposition of the basic principles of astronautics, ". . .the science and technology of the placement and control of manned and unmanned objects in space." The book is based upon a course instituted by Professor Berman in 1958 and develops most of the topics covered from first principles.

Those possessing a background in modern physics, such as Berman himself, will find the book to be of particular advantage. Chapter One discusses the principles of astronomy and divides the field into cosmogony, astrophysics, and celestial mechanics. Here as in the rest of the text the areas of practical astronomy, navigational astronomy, and the burgeoning field of meteoritics have been neglected. These disciplines enrich the background of astronautics and should probably have been treated, if only in a cursory fashion. Aside from these omissions, the introductory material is clear and presented in an exceedingly interesting fashion.

The physics of the solar system chapter is equally well done and a careful reading shows that the author has made every effort to include the most up-to-date data and astrodynamic constants available at the time of publication—a very painstaking task indeed. If any criticism is to be leveled at the subsequent chapter on the foundations of mechanics, it would be that it is a bit too clear, too elementary. It would, of course, serve as a valuable review but hopefully the reader would have had some prior exposure to such subjects as Newton's laws. On the other hand the discussion of accelerated reference coordinates is very relevant and deserves the elaboration given this chapter. In dealing with potential and kinetic energy the *vis-viva* equation is developed and utilized in some simple illustrative examples. It is felt, however, that the full power of this exceedingly useful little equation is mitigated by the numerous different ways in which it is formulated.

The subject of the dynamics of flight is also effectively presented but suffers somewhat from the overly emphasized analogy with potential barriers so useful to atomic physics and the lack of discussions of orbit determination and orbit prediction techniques. Another drawback can be found in the otherwise laudable attempt to utilize exclusively the metric system of units. Accuracy, simplicity of formula, and historical precedent all recommend the dimensionless set of astrodynamic units (e.g., the Earth radius, a.u., speed at unit distance, etc.). The chapter on dynamics of flight, as well as many of the other chapters, is well illustrated by numerical examples—such examples are indispensable tools in class instruction.

Transfer orbits and space navigation are clearly developed, again with a number of useful numerical examples. The only two criticisms that apply here are that "A fundamental rule of astronautics" (p. 179) is not clearly shown to be a logical consequence of the variation of parameters, and that the use of differential correction is not discussed. The chapter on orbit perturbation is well illustrated and gives the student a clear intuition into the physics of "what's going on."

Finally, a chapter on propulsion dynamics concludes the book. This chapter is particularly well done and shows that the author is knowledgeable in almost all areas of conventional and unconventional propulsion. Our only regret is that so interesting and well done a topic was not extended even further

From almost every point of view *The Physical Principles of Astronautics* is a fine book and should be recommended as an introductory text to most fields of astronautics. In fact as Professor Berman states in the preface, "This book . . . will serve to lay the groundwork at an intermediate level, for works in astrodynamics . . . ," and presumably also for works in propulsion, spacecraft structures, aerodynamics, communications, and the other divisions of astronautics.

ROBERT M. L. BAKER, JR.
Department of Engineering
University of California
Los Angeles, and
Lockheed California Company,
Astrodynamic Research Center

MUrray Hill 2-3383

AMERICAN ASTRONAUTICAL SOCIETY

516 FIFTH AVENUE, NEW YORK 36, NEW YORK

OFFICE RECORD

REC'D.	
REC'D.	
ACK.	
R.R.	
ELECT.	
GRADE	
DATE	
NO.	

APPLICATION FOR MEMBERSHIP

) THE MEMBERSHIP COMMITTEE OF THE AMERICAN ASTRONAUTICAL
)CIETY:

ıereby apply for membership in the American Astronautical Society,
d do agree to conform to the Constitution and By-Laws of the Society.
e information given on this application is true and correct to the best
my knowledge.

ATE_____ SIGNATURE IN FULL_____
(Sign with pen—Do not use initials)

BIOGRAPHICAL RECORD

(Please fill in all information by hand lettering or typewriter)

ame_____ Age_____
(First) (Middle) (Last)

] Home Address_____

] Business Address_____
(Check mailing address)

ome Tel. No._____ Bus. Tel. No._____

ate of Birth_____ Place_____

ex_____ Country of Citizenship_____

Company or Organization with which you are connected_____

_____

Your Title or Position_____

Nature of your Duties, or your Profession or Occupation_____

_____

Name of Technical or Scientific Organizations to which you belong. (State Grade of Membership)_____

_____

_____

f you are a Licensed or Registered Engineer or other Professional Specialist, give Title of such License and

State in which Registered_____

_____

EDUCATIONAL RECORD

College or University	Dates of Attendance	Major Subject	Degree Rec'd & D
Other			

Awards, Fellowships, Scholarships or other Honors Received (Give Dates and Citations)

State Titles of Scientific Papers and When and Where They Were Presented and/or Published

REFERENCES

List at least four references familiar with your activities. Complete addresses of non-members must be su plied. However, references need not be A.A.S. members.

1.
2.
3.
4.
5.

Grade of Membership for Which You Apply

PROFESSIONAL RECORD

(Needs only to be completed by Applicants for the Following Grades: Affiliate Member, Member, or Ass ciate Fellow.)

Give a brief outline of your Industrial, Academic or Related Experience. Use additional sheets necessary.

Date, From - To (year only)	Name of Affiliation and Description of Duties

Format of Technical Papers for THE JOURNAL OF THE ASTRONAUTICAL SCIENCES

The Editors will appreciate the cooperation of authors in using the following directions for the preparation of manuscripts. These directions have been compiled with a view toward eliminating unnecessary correspondence, avoiding the return of papers for changes, and reducing the charges made for "author's corrections."

Manuscripts

Papers should be submitted in original typewriting (if possible) on one side only of white paper sheets, and should be double or triple spaced with wide margins. However, good quality reproduced copies (e.g. multilith) are acceptable. An additional copy of the paper will facilitate review.

Company Reports

The paper should not be merely a company report. If such a report is to be used as the basis for the paper, appropriate changes should be made in the title page. Lists of figures, tables of contents, and distribution lists should all be deleted.

Titles

The title should be brief, but express adequately the subject of the paper. A footnote reference to the title should indicate any meeting at which the paper has been presented. The name and initials of the author should be written as he prefers; all titles and degrees or honors will be omitted. The name of the organization with which the author is associated should be given in a separate line to follow his name.

Abstracts

An abstract should be provided, preceding the introduction, covering contents of the paper. It should not exceed 200 words.

Headings

The paper can be divided into principal sections as appropriate. Headings or paragraphs are not numbered.

Illustrations

Drawings should be made with black India ink on white paper or tracing cloth, and should be at least double the desired size of the cut. Each figure number should be marked with soft pencil in the margin or on the back of the drawing. The width of the lines of such drawings and the size of the lettering must allow for the necessary reduction. Reproducible glossy photographs are acceptable. However, drawings which are unsuitable for reproduction will be returned to the author for redrawing. Legends accompanying the drawings should be typewritten on a separate sheet, properly identified.

Security Clearance

Authors are responsible for the security clearance by an appropriate agency of the material contained in the papers.

Mathematical Work

As far as possible, formulas should be typewritten. Greek letters and other symbols not available on the typewriter should be carefully inserted in ink. Each such symbol should be identified unambiguously the first time it appears. The distinction between capital and lower-case letters should be clearly shown. Avoid confusion between zero (0) and the letter *O*; between the numeral (1), the letter l, and the prime (′); between alpha and *a*, kappa and *k*, mu and *u*, nu and *v*, eta and *n*.

The level of subscripts and exponents, should be clearly indicated. Vectors will be set in bold face type. Authors should indicate this in their manuscripts by a wavy underscore.

Greek Alphabet

A	α	alpha	(a)	N	ν	nu	(n)
B	β	beta	(b)	Ξ	ξ	xi	(x)
Γ	γ	gamma	(g)	O	o	omicron	(ŏ)
Δ	δ	delta	(d)	Π	π	pi	(p)
E	ϵ	epsilon	(ĕ)	P	ρ	rho	(r)
Z	ζ	zeta	(z)	Σ	σs	sigma	(s)
H	η	eta	(ē)	T	τ	tau	(t)
Θ	θ	theta	(th)	Υ	υ	upsilon	(u)
I	ι	iota	(i)	Φ	$\phi\varphi$	phi	(ph)
K	κ	kappa	(k)	X	χ	chi	(ch)
Λ	λ	lambda	(l)	Ψ	ψ	psi	(ps)
M	μ	mu	(m)	Ω	ω	omega	(ō)

The Orbital Elements

a = semimajor axis
e = eccentricity
Ω = longitude of the ascending node
i = inclination
ω = argument of perifocus
T = time of perifocal passage

Complicated exponents and subscripts should be avoided when possible to represent by a special symbol.

Fractions in the body of the text and fractions occurring in the numerators or denominators of fractions should be written with the solidus as follows:

$$\frac{\cos(\pi x/2b)}{\cos(\pi\alpha/2b)}. \qquad (1)$$

The intended grouping of handwritten formulas can be made clear by slight variations in spacing, but this procedure is not acceptable in printed formulas. To avoid misunderstanding, the order of symbols should therefore be carefully considered. Thus:

$(a + bx)\cos t$ is preferable to $\cos t\,(a + bx)$.

In handwritten formulas the size of the braces, brackets, and parentheses can vary more widely than in print. Particular attention should be paid to the proper use of braces, brackets, and parentheses (which should be used in this order). Thus:

$$\{[a + (b + cx)^n]\cos ky\}^2 \qquad (2)$$

is required rather than $((a + (b + cx)^n)\cos ky)^2$.

Equations are centered, numbered, punctuated as complete sentences; and referred to in text as (2).

Bibliography

References should be grouped together in a bibliography at the end of the manuscript. References to the bibliography should be made by numerals between square brackets [4].

The following examples show the approved arrangements:

for books—[1] BAKER, R. M. L., Jr. and MAKEMSON, M. W., *An Introduction to Astrodynamics*, Academic Press, New York, 1st ed., 1960.

for periodicals—[2] LAMORE, LEWIS, "Celestial Observations for Space Navigation," J. Astronaut. Sci., 6 (1959), 1–10.